electronics

MDPI

Article

MBSDet: A Novel Method for Marine Object Detection in Aerial Imagery with Complex Background Suppression

Gaofan Ji [1,2], Li Fan [1,2,3,*] and Chuanxiang Li [4,*]

[1] School of Engineering, Huzhou University, Huzhou 313000, China; 2022388610@stu.zjhu.edu.cn
[2] Huzhou Institute of Zhejiang University, Huzhou 313000, China
[3] College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
[4] Xi'an Research Institute of High-Tech, Xi'an, Shaanxi 710025, China
* Correspondence: fanli77@zju.edu.cn (L.F.); lichuanxiang@zju.edu.cn (C.L.)

Abstract: Detecting maritime ship targets in complex aerial images is significantly challenging due to the relatively small size of the ship compared to the image's predominant background. Current methodologies used to identify maritime targets in aerial imagery are susceptible to disruption from intricate backgrounds, leading to misidentification, overlooked detection, and inaccurate target localization. To address these challenges, we proposed the maritime background suppression network (MBSDet), which facilitates the reliable detection of multi-scale targets in intricate backgrounds. Firstly, we developed a background suppression (BS) module that integrates the Transformer's context-aware capabilities with local attention features to enhance the model's ability to identify multi-scale objects. Secondly, we present the multidimensional feature enrichment (MFE) module, which employs feature pyramid-based rearrangement modules and dynamic upsampling operators instead of conventional upsampling methods to enhance feature richness while attenuating background features. The mean average precision (mAP) values can reach 90.54% and 88.89% on HRSC2016 and DOTA v1.0 datasets, respectively. The experimental results indicate that MBSDet exhibits outstanding detection and localization accuracy for objects in difficult background situations.

Keywords: marine object detection; aerial image; multi-scale target; background feature suppression; transformer; attention

check for updates

Citation: Ji, G.; Fan, L.; Li, C. MBSDet: A Novel Method for Marine Object Detection in Aerial Imagery with Complex Background Suppression. *Electronics* **2024**, *13*, 4764. https://doi.org/10.3390/electronics13234764

Academic Editor: Duc Thanh Nguyen

Received: 16 October 2024
Revised: 27 November 2024
Accepted: 30 November 2024
Published: 2 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ship recognition in aerial images, due to its extensive coverage and distinct high-altitude perspective, has been widely utilized in various fields such as coastal patrol [1,2], marine environmental assessment [3], and maritime rescue operations [4,5]. Conventional maritime object identification techniques frequently rely on human-designed feature extraction from pictures. The traits are subsequently integrated to categorize things. Manually crafted features are also unstable, intricate, and exhibit limited generalizability. These conventional approaches are constrained in their practical uses. Consequently, the emphasis of studies on target identification methodologies has progressively transitioned toward the utilization of deep learning techniques.

Convolutional neural networks (CNNs) for object detection are categorized into the following two-stage methods: Cascade R-CNN [6], which uses region proposals and subsequent classification and regression, and one-stage approaches, such as YOLO [7], SSD [8], and MobileNet [9], which offer end-to-end architectures for direct classification and localization. Enhancements to these methods aim to increase precision while maintaining rapid detection capabilities. Object detection methodologies are also classified based on the use of predefined bounding boxes into anchor-based and anchor-free techniques [10]. Traditionally, two-stage methods like Cascade R-CNN are anchor-based and

rely on the candidate's prior bounding boxes, while one-stage methods have been predominantly anchor-based. However, recent developments have introduced anchor-free techniques [11–14] that achieve comparable accuracy. Object detection algorithms often utilize anchor points to predict target locations, with examples including RetinaNet [15] and YOLOv11 [16]. ReDet [17], as an advanced two-stage network, enhances accuracy in aerial object detection through its rotation-equivariant features and rotation-invariant RoI Align, yet it introduces increased computational complexity, higher training difficulty, and a more intricate implementation process. The OAB [18] simplifies detection with a single anchor point, reducing hyperparameter tuning, but struggles with complex backgrounds of objects. YOLOv11 [16] addresses this by using nine unique anchors for various object sizes, yet challenges remain in detecting objects like ships of different sizes and orientations in aerial imagery datasets such as HRSC2016 [19]. FCOS [20] improves network efficiency through a per-pixel prediction approach, but for large-stride network feature maps, it may miss some objects. The KLD method [21], while enhancing precision and scale invariance in remote sensing image object detection, faces challenges such as high computational complexity, strong parameter sensitivity, great implementation difficulty, and unverified generalization capability. The phase-shifting coder (PSC) [22] enhances detection accuracy and robustness by addressing boundary discontinuities and square-like issues, but it may face challenges in computational complexity and model generalizability. R3Det [23] enhances rotating object detection accuracy and efficiency with feature refinement and an approximate SkewIoU loss, but it has high computational complexity and requires meticulous hyperparameter tuning. S2A-Net [24] improves the accuracy of aerial image detection through feature alignment and oriented detection modules, but it faces challenges in model complexity and computational resource consumption [25,26]. We aim to design an end-to-end aerial image object detection model that exhibits high robustness to targets with complex backgrounds. Compared to the aforementioned networks, RTMDet [27], as a state-of-the-art model, demonstrates excellent detection performance and efficiency on both the DOTA dataset and the HRSC2016 dataset. However, it exhibits weaker robustness to targets with complex background information. Therefore, this paper conducts research based on RTMDet.

Background suppression of object detection is needed for numerous applications, including aerial object detection, which entails recognizing images obtained from cameras or sensors external to large or small ships. This technology is crucial for applications like surveillance [4], precision agriculture, and urban administration [28,29]. Genin et al. [30] introduced an innovative background suppression approach for small object detection; yet, it may inadequately address the intricacies of changing backgrounds, thereby compromising the algorithm's resilience in real-world applications. Kim et al. [31] enhanced the signal-to-noise ratio to mitigate background interference; nonetheless, their approach may be constrained in intricate and varied environments, thereby diminishing detection accuracy. Kushwaha et al. [32] proposed a framework for dynamic background modeling and shadow suppression that, while effective, may necessitate substantial computational resources, hence restricting its use in resource-limited settings. Kushwaha et al. [33] accomplished dynamic object detection via shadow suppression; nevertheless, they may encounter difficulties with aerial photography targets. Zheng et al. [34] created a road defect detection network that proficiently isolates and mitigates background information, but it may encounter challenges in managing intricate road conditions. Pan et al. [35] proposed a background suppression pyramid network to mitigate background information; however, it may encounter difficulties in addressing multi-scale targets within intricate backdrop images.

The RTMDet framework [27], which serves as a benchmark, achieves remarkable accuracy in object detection by employing deep convolutional networks with large kernel sizes and a flexible soft label assignment technique. However, it falls short in improving detection accuracy for objects of varying sizes. To begin with, the multiple layers in the backbone network repeatedly perform downsampling during the feature map extraction

process, leading to a decrease in resolution and, as a consequence, a loss of detail necessary for detecting multi-scale objects. In addition, the deep feature maps derived from the backbone network have low resolution and large receptive fields, which are not ideal for detecting smaller objects. Moreover, tiny objects are particularly vulnerable to interference within limited spatial areas. Building upon the previously discussed tiny object identification techniques, our MBSDet model incorporates two novel modules: the multidimensional feature enrichment (MFE) module and the background suppression (BS) module. We initially substituted conventional upsampling with a dynamic learning methodology, enabling the model to autonomously determine the locations of the upsampling samples. Subsequently, we developed the BS module, which utilizes the long-range feature correlation abilities of Transformers in conjunction with local attention processes to amplify focus on multi-scale objects and diminish background characteristics [36,37]. Li et al. [38] introduced an innovative Transformer module that refined Transformer architecture in visual tasks by leveraging contextual information among input keys to inform the learning of dynamic attention matrices, thus augmenting visual representation efficacy. Chen et al. [39] developed a sequential Transformer for autoregressive pixel prediction, attaining performance on par with CNNs in image classification tasks. Vision Transformer (ViT) is another visual Transformer model that utilizes pure Transformers directly on sequences of image patches for comprehensive image classification. Dosovitskiy et al. [40] recently presented a novel approach to image recognition that attained superior performance across many benchmarks. Nicolas Carion et al. [41] utilized Transformer architecture to directly forecast sets of items, circumventing conventional region candidates or anchors. Zhu et al. [42] presented Deformable DETR, a deformable Transformer framework intended for comprehensive object detection. Zhang et al. [43] formulated a Transformer-based YOLO architecture for object-detecting applications. Li et al. [44] introduced the MViTv2 network architecture, which features an advanced vision Transformer design. Consequently, we present an aerial ship object recognition approach, MBSDet, which exhibits superior recognition accuracy on the HRSC2016 and DOTA v1.0 [45] datasets.

This study's primary contributions are summarized as follows:

1. MBSDet proposal: To address the limitations encountered by current object recognition techniques for aerial ship imagery, we present the MBSDet method. This technique mitigates intricate background elements and enhances the detection precision of multi-scale targets.

2. Introduction of the multidimensional feature enrichment (MFE) module: This module facilitates more adaptable sampling positions while preserving resolution. When combined with the pyramid network, it incorporates contemporary features with lower-dimensional attributes to derive richer foreground features. Utilizing an offset matrix, it adaptively acquires the offsets of sampling points, allowing the network to identify more suitable upsampling features throughout the learning process.

3. Development of the background suppression (BS) module: This module integrates the long-range dependency association and local attention features of the Transformer to obtain more specific knowledge about objects while minimizing background information. By segmenting into various receptive fields, it differentiates between foreground and background features.

Experimental findings indicate that MBSDet attains superior accuracy on the HRSC2016 and DOTA v1.0 datasets, showcasing its significant capacity to mitigate background characteristics in intricate situations.

2. Methods

In this section, we introduce the overall architecture of our MBSDet model. Subsequently, we delve into the specifics of the background suppression (BS) module and the multidimensional feature enrichment (MFE) module.

2.1. Overview of MBSDet

Considering that targets in aerial photographs frequently occupy small areas, while the backgrounds constitute the majority of the images, the impact of background features on target characteristics must not be overlooked. To mitigate background characteristics, we implemented our BS module subsequent to the *C*3, *C*4, and *C*5 phases of the backbone network, as seen in Figure 1. In this context, "C_i" represents the feature map after the convolution operation, while "P_i" denotes the feature map after the upsampling process. This seeks to mitigate background information by utilizing the multi-scale attention and contextual information correlation features of the Transformer. In the feature fusion phase, we substitute the conventional upsampling module with our MFE module, which more effectively retains details throughout the upsampling of feature maps.

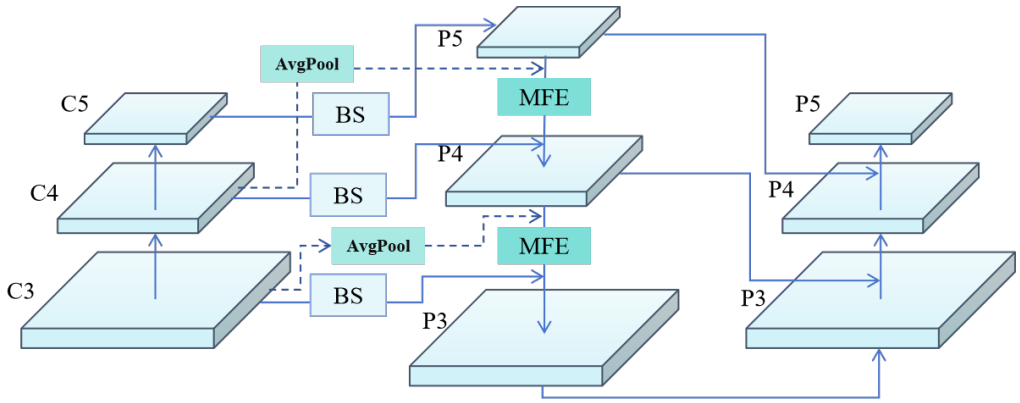

Figure 1. An overview of our MBSDet model. BS module: background suppression module. MFE module: multidimensional feature enrichment module.

2.2. BS Module

RTMDet has demonstrated the efficiency of its backbone network through extensive experiments; therefore, we continue to adopt CSPNeXt as the backbone network and make improvements to the neck part. As shown in Figure 2, the Transformer module's self-attention mechanism can effectively connect contextual information and local attention characteristics when detecting objects in aerial images. Firstly, the image is divided into patches of size $p \times p$. Then, each patch is subjected to positional encoding, which is input into the encoder–decoder structure. Finally, the detection results are output through a multilayer perceptron (MLP). This approach allows the model to comprehend the interdependencies across the entire image, thereby improving its grasp of the overall structure and context. Furthermore, the Transformer module's multi-head attention mechanism augments the model's capacity to detect features at various scales and orientations. We used these good qualities to create the BS module, which, as shown in Figure 3, takes the input feature map and processes it through three separate branches to improve key information and reduce background noise. The first branch focuses on local features using a 5×5 patch size for window division, followed by feature extraction through the Transformer module. The second branch employs a larger 25×25 patch size to capture broader background features, while the third branch utilizes depthwise separable convolutions. Finally, we merge the outputs from these three branches to produce the total output feature.

Assuming the input is $X \in \mathbb{R}^{B \times C \times H \times W}$, with B denoting the batch size, C denoting the number of channels, and H and W denoting the height and width of the image. Branch 1 and Branch 2 can be represented as follows:

$$U_i(T_i(P_i(X))) = Transpose(T_i(P_i(X)), kernelsize = P_i, stride = P_i) \quad i = 1, 2 \tag{1}$$

Here, P_i denotes the patch size, T_i indicates the feature extraction performed by the Transformer on the patches i, and U_i denotes the transformation of the image back to its original format after processing by the Transformer. Branch 3 can be represented as follows:

$$D_3(X) = DSConv(X, kernel_size = 3, padding = 1, groups = C) \quad (2)$$

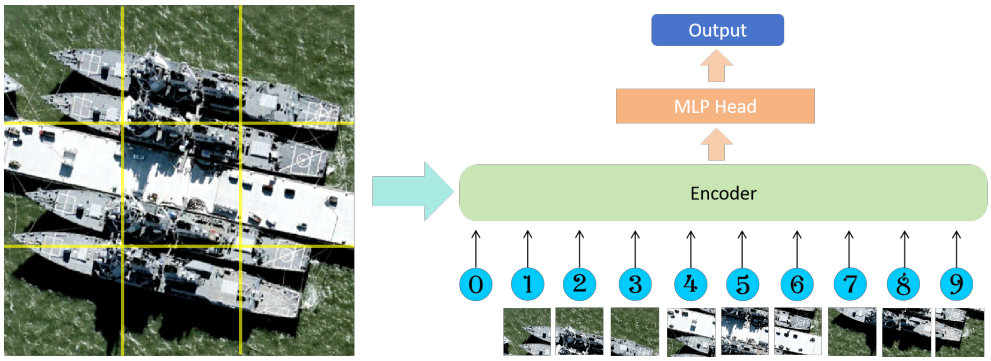

Figure 2. Flow chart of the Transformer module.

Here, DSConv is composed of one 3×3 convolution and one 1×1 convolution. The final output formula can be represented as follows:

$$O = U_1(T_1(P_1(X))) + U_2(T_2(P_2(X))) + D_3(X) \quad (3)$$

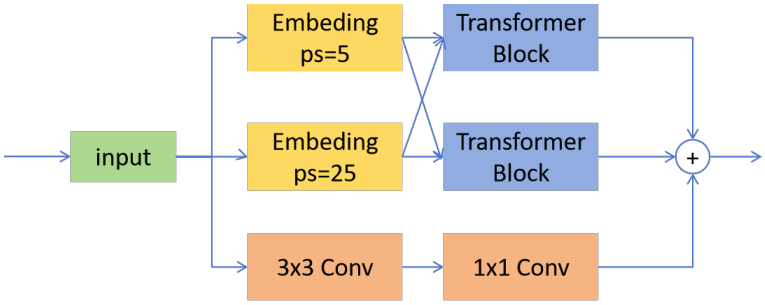

Figure 3. The background suppression module. ps = 5×5, 25×25 refers to the dimensions of each block after the image is partitioned. 3×3 and 1×1 refer to the kernel size of the convolutional filters being 3×3 and 1×1, respectively.

2.3. MFE Module

Generally, the neck feature fusion network resamples feature maps from various resolutions to a uniform resolution prior to concatenation. Many contemporary approaches for target detection in aerial datasets rely on the original nearest neighbor interpolation technique found in feature pyramid networks (FPNs). This method fills in the feature maps with pixels that are next to each other and only looks at the subpixel neighborhood. However, this method often loses a lot of feature information. 'DyUpsample' [46] reconsiders the upsampling process by introducing sampling offsets, allowing the network to learn the upsampling features itself. However, this method only considers the feature maps of the current dimension and does not handle multi-scale targets well.

Therefore, we present the MFE module, as depicted in Figure 4. Before upsampling, we perform subtraction operations between the feature maps of the current dimension and those of the previous dimension, followed by activation using the ReLU function. The rationale behind this is that aerial images predominantly consist of the background; hence, subtracting the average-pooled feature map of the previous dimension from the feature map of the current dimension helps to effectively suppress the impact of background

features. Assuming the input is $X \in \mathbb{R}^{B\times C\times H\times W}$, $LX \in \mathbb{R}^{B\times C\times H'\times W'}$ represents the low-dimensional feature map, and the subtracted feature graph $X_{\text{processed}}$ is calculated first as follows:

$$X_{\text{processed}} = X - Pooling(LX) \tag{4}$$

The formula for calculating the offset can be expressed as follows:

$$O_{\text{offset}} = \text{Conv2d}\left(X, out_channels = 2 \times groups \times scale^2, kernelsize = 1\right) \tag{5}$$

$$O = O_{offset} \times 0.25 + init_pos \tag{6}$$

$$O = ReLU(BatchNorm2d(Conv2d(O, kernel_size = 3, padding = 1)) + O) \tag{7}$$

Finally, we upsample the processed feature graph, and the formula is as follows:

$$\text{MFE}(X, L_X) = S\left(X_{\text{processed}}, O_{\text{res}}\right) \tag{8}$$

As shown in Figure 1, during the upsampling processes of P4 and P5, the average-pooled feature maps of C3 and C4 are first subtracted, respectively, before upsampling occurs. We then apply dynamic upsampling to the subtracted feature maps. Thus, our MFE module can effectively leverage multidimensional information to suppress background features and enrich foreground features.

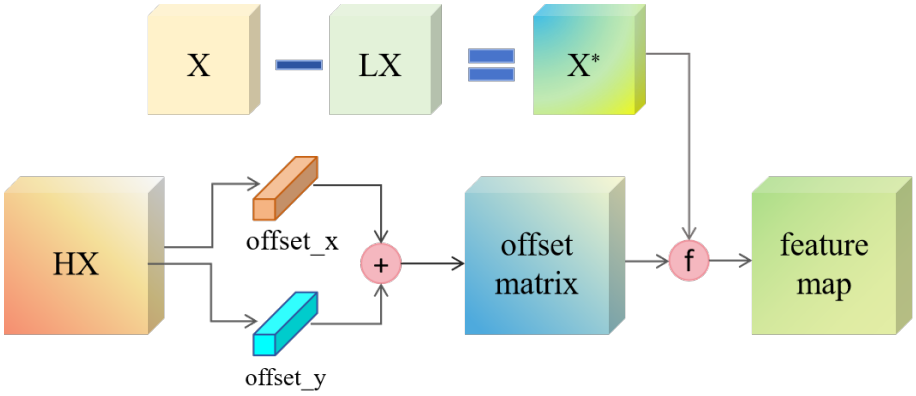

Figure 4. The framework of the multidimensional feature enrichment module. X, HX, LX, and X* represent the input feature map, the high-dimensional feature map, the low-dimensional feature map, and the calculated feature map, respectively.

3. Results

This study utilized the HRSC2016 and DOTA v1.0 datasets to assess the proposed MBSDet in conjunction with other comparative methodologies.

3.1. Datasets and Performance Metrics

HRSC2016 dataset: The HRSC2016 dataset [19] is a publicly available maritime ship detection dataset. It consists of 1061 very high-resolution remote sensing images that Northwestern Polytechnical University collected from Google Earth. The images vary in size from 300 × 300 to 1500 × 900 pixels. The HRSC2016 dataset contains 2976 instances of ship objects from across 33 categories. We used the rotated bounding box annotations from the HRSC2016 dataset to evaluate various methods.

DOTA v1.0 dataset: The DOTA v1.0 dataset is sourced from images acquired via Google Earth, the GF-2 satellite, the JL-1 satellite, and aerial photography provided by CycloMedia B.V. It comprises photos of diverse resolutions, generally varying from 800 pixels by 800 pixels to 4000 pixels by 4000 pixels. The collection encompasses 15 prevalent item categories, including planes, ships, and storage tanks, totaling 188,282 object instances,

each labeled with arbitrary quadrilaterals. The rotated bounding box annotations from the DOTA v1.0 dataset were utilized to assess various methodologies.

Evaluation metrics: To evaluate the detection accuracy of various approaches for remote sensing targets, we used the COCO evaluation criteria. The following metrics are included: AP50, which represents the mean average precision at an Intersection over Union threshold of 0.5 across all categories, and mAP, which stands for the mean average precision calculated over IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05. The formula is as follows:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{9}$$

3.2. Implementation Details

All experiments in this study were conducted using PyTorch on a single NVIDIA GeForce RTX 3090 GPU. The HRSC2016 and DOTA v1.0 datasets were divided in a 3:1:2 ratio. The HRSC2016 dataset comprises 538 images for training, 178 images for validation, and 356 images for testing. The DOTA v1.0 dataset was segmented into 1024 × 1024 images, comprising a total of 31,881 images, with 15,734 allocated for the training set, 5313 for the validation set, and the original 10,834 designated for the test set. Through reviewing the relevant literature, it has been determined that setting the training parameters to a total of 300 epochs, a batch size of 8, and an initial learning rate of 0.002 yields optimal results; therefore, we also set the training parameters accordingly.

3.3. Comparison of Experiments

To verify the performance of MBSDet in aerial image object recognition, we carried out comparative analyses with several advanced approaches on the HRSC2016 and DOTA v1.0 datasets. The methodologies being compared include H2RBox-v2 [47], ReDet [17], Rotated-FCOS [20], KLD [21], PSC [22], S2ANet [24], YOLOv11 [16], R3Det [23], and RetinaNet. The quantitative comparison results for the HRSC2016 and DOTA v1.0 datasets are presented in Tables 1 and 2, with the better results highlighted in green font. MBSDet achieved the highest mean average precision (mAP) on both the HRSC2016 and DOTA v1.0 datasets, as well as at the 0.5 Intersection over Union (IoU) thresholds.

Results on the HRSC2016 dataset: Compared to conventional one- and two-stage object detection techniques, MBSDet produced target detection outcomes that were more accurate, as shown in Table 1. MBSDet improved the mAP by 5.29% and the AP50 by 5.44% compared to the robust baseline RTMDet. ReDet, an advanced two-stage algorithm, achieved extremely high precision, as the table shows; however, our MBSDet had fewer parameters.

Results on the DOTA v1.0 dataset: Correspondingly, MBSDet attained the highest object detection performance among the evaluated methods on the DOTA v1.0 dataset, as illustrated in Table 2. When compared with the robust baseline RTMDet, MBSDet increased mAP by 4.47%, reaching an AP50 of 89.20%. Quantitative results indicate that MBSDet can effectively manage complex background features.

Table 1. Results on the HRSC2016 dataset. The best results are shown in green.

Model	mAP	AP50	Gflops	Params (M)
H2RBox-v2	57.82	89.80	206.3	31.89
ReDet	90.42	90.40	-	121
Rotated_fcos	84.47	84.50	206.2	31.89
KLD	83.91	84.50	209.58	36.13
PSC	83.51	83.50	210.94	36.19
Retina	82.80	84.88	209.58	36.19
RTMDet	85.25	85.16	20.45	4.87
MBSDet (ours)	90.54	90.60	22.12	6.51

Figure 5 illustrates the feature visualization results of different models on the same image; the red areas indicate higher feature values corresponding to the target features, and the blue areas indicate lower feature values corresponding to the background features. The red areas have a greater impact on the prediction results. The desired feature map visualization displays the target area in clear red and other areas in blue, which can lead to more accurate prediction results. Figure 5a shows the feature visualization result of the RetinaNet model, where the entire feature map is mostly red, indicating that features are not concentrated in the ship area, suggesting that background features such as the sea surface will affect the detection results. Figure 5b shows the feature visualization result of the H2RBox-v2 model, where the ship area is predominantly red, and other parts are blue, which may result in more accurate predictions. Figure 5c shows the feature map visualization result of the baseline RTMDet model, from which it can be seen that most areas outside the ship are red, indicating that the background will have a significant impact on the model's prediction results. Figure 5d shows the feature visualization result of our MBSDet model, from which it can be seen that the area around the ship is red and the aspect ratio also conforms well to the characteristics of the ship target, with other areas being blue. Compared with Figure 5a and Figure 5c, our model displays the ship area in red and other areas in blue, indicating that our model can effectively suppress background information. Compared with Figure 5b, our model not only highlights the ship area but also maintains an aspect ratio that mirrors that of the ship, indicating that our model will be more precise in predicting the position of the ship. It should be noted that the slight offset of the red areas relative to the ship in all figures is due to the scaling of the feature map, but this does not affect the analysis of the results. The comparison of quantitative data and visualizations demonstrates that our MBSDet effectively suppresses background characteristics, thereby improving its resilience in detecting multi-scale targets.

Table 2. Results on the DOTA v1.0 dataset. The best results are shown in green.

Model	mAP	AP50
H2RBox-v2	81.40	81.40
ReDet	84.80	84.80
Oriented R-CNN	73.40	74.60
Rotated_fcos	74.89	75.40
KLD	80.50	80.50
PSC	75.32	75.60
Retina	76.60	76.89
S2ANet	81.73	81.70
R3Det	75.30	75.30
YOLOv11-s	74.06	51.04
YOLOv11-m	76.00	53.07
YOLOv11-l	76.40	53.80
RTM-s	81.98	82.73
RTM-m	82.66	83.24
RTM-l	84.42	85.85
MBSDet-tiny	85.53	85.50
MBSDet-s	87.20	87.20
MBSDet-m	88.88	88.90
MBSDet-l	88.89	89.20

According to the differences in the depth and width of the network, s, m, and l represent the small, medium, and large versions of the model, respectively.

Figure 6 presents the detection results of several models on the HRSC2016 dataset. Specifically, column (a) represents the actual or correct results, column (b) highlights the detection results of our MBSDet model, column (c) displays the detection results of the baseline RTMDet model, and column (d) illustrates the detection results of the typical two-stage detection network, ReDet. In columns (b), (c), and (d), the red solid-line rectangles indicate the positions of ships predicted by different models. Compared to the correct

results in column (a), we mark the undetected ships, such as the first and second rows in the third and fourth columns, with green-dashed rectangles. For the rectangles with incorrectly predicted categories, such as the third column in the third row, the third and fourth columns in the fourth row, and the third column in the fifth row, we mark them with cyan-dashed rectangles. As for the results with incorrectly predicted positions, such as the third column in the sixth row, where the rectangle includes parts that do not belong to the ship, we mark them with cyan-dashed rectangles.

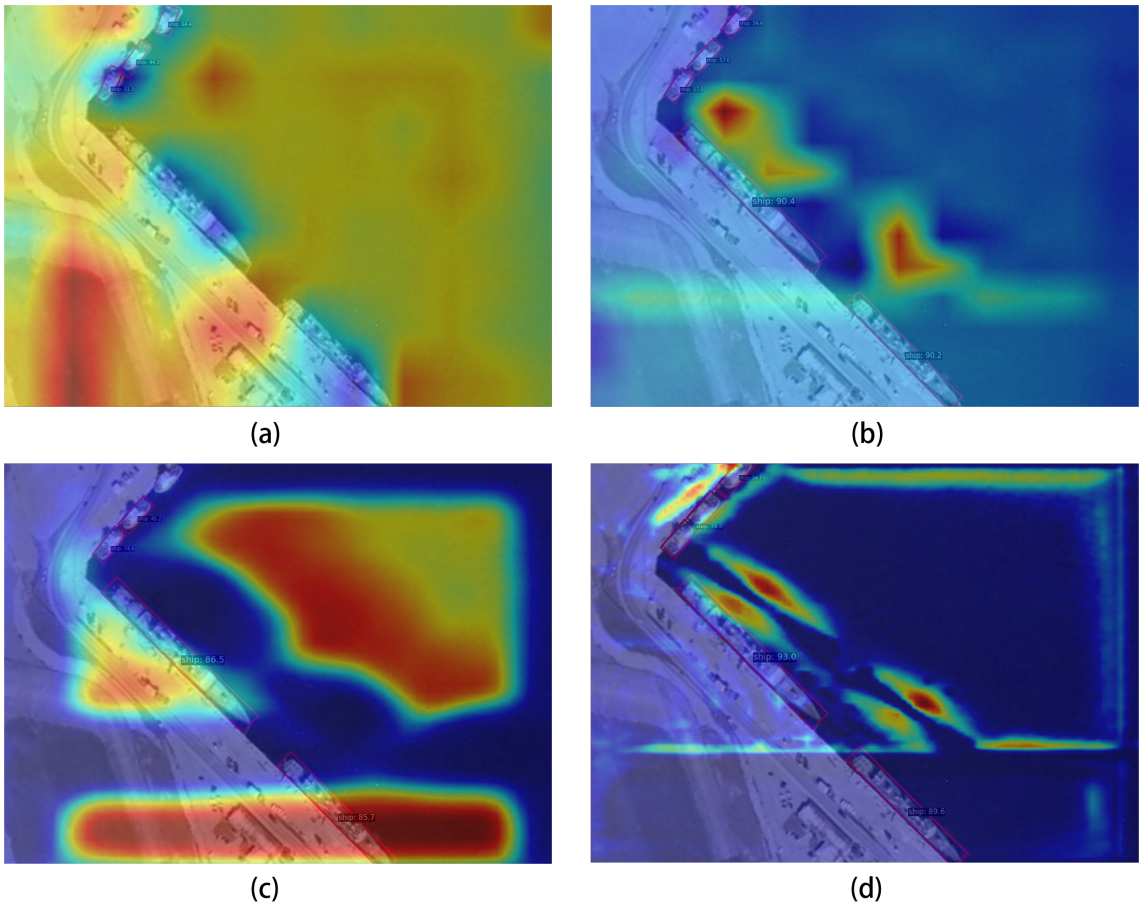

Figure 5. Feature visualization outcomes of various methods. (**a**) RetinaNet. (**b**) H2RBox-v2. (**c**) RTMDet. (**d**) MBSDet.

Figure 7 displays the detection results of several models on the DOTA v1.0 dataset. In this figure, column (a) represents the actual or correct results, column (b) presents the detection results of our MBSDet model, column (c) shows the detection results of the baseline RTMDet model, and column (d) illustrates the detection results of the typical two-stage detection network, ReDet. The red rectangles in the figure represent ships, yellow rectangles represent harbors, purple rectangles represent small vehicles, and white rectangles represent planes. Compared to the correct results in column (a), we mark the undetected targets with green-dashed rectangles. These include the third and fourth columns of the first row, where harbors were not detected, and the third and fourth columns of the third row, where ships were not detected. For rectangles with incorrect predicted categories or positions, we mark them with cyan-dashed rectangles. For example, in the second row's third and fourth columns, harbors were incorrectly predicted as ships; in the third row's fourth column, objects on the ship were incorrectly predicted as a harbor; in the fourth row's third and fourth columns, the background was incorrectly predicted as a harbor; in the fifth row's third column, a harbor was incorrectly predicted as a ship; and in

the sixth row's third and fourth columns, ground markings were incorrectly predicted as planes.

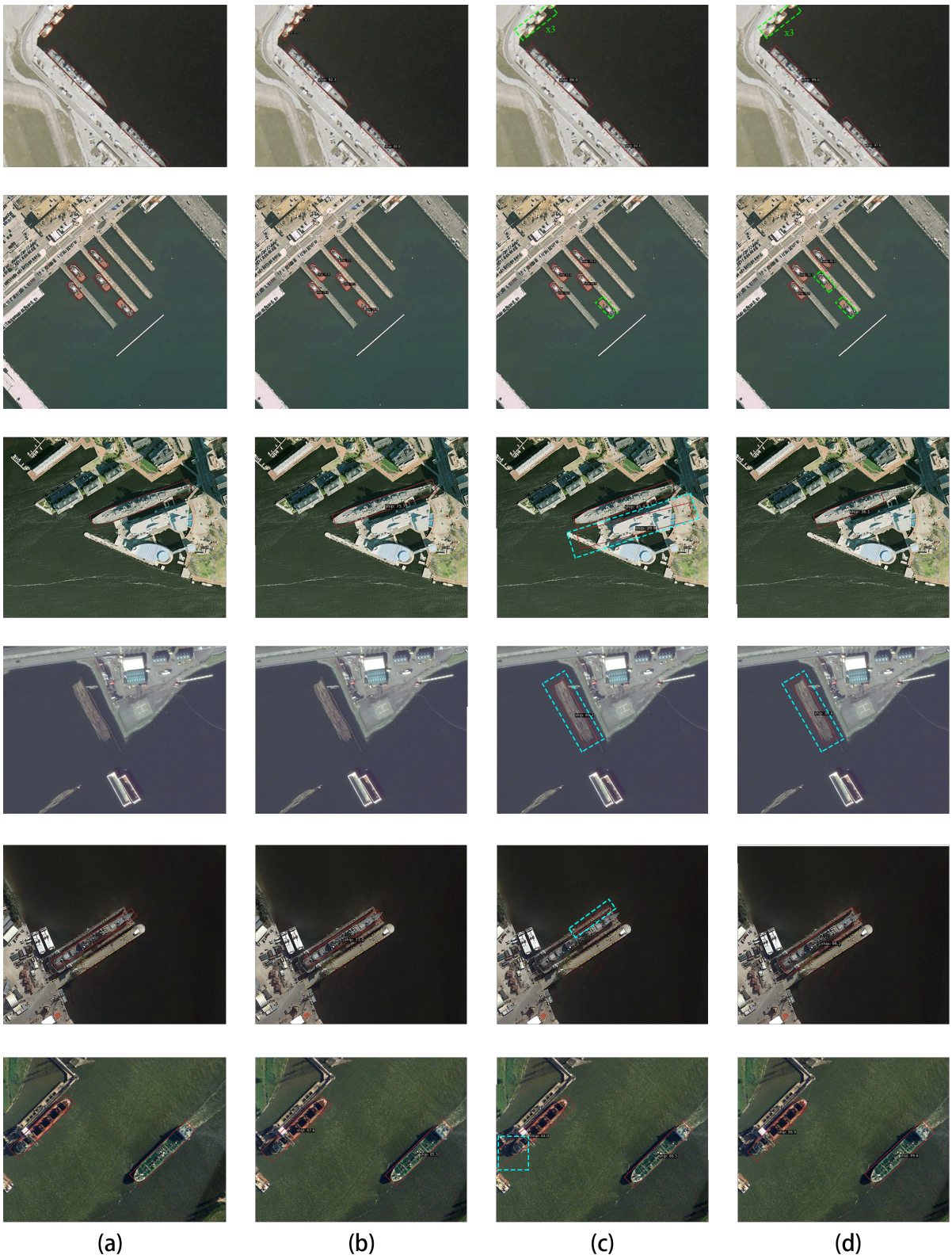

Figure 6. Results from various methods on the HRSC2016 dataset. (**a**) Ground truth. (**b**) MBSDet. (**c**) RTMDet. (**d**) ReDet. The dashed-green box indicates missed detection, and the dashed-cyan box indicates false detection. Solid line boxes represent the actual or predicted results; red represents the ship.

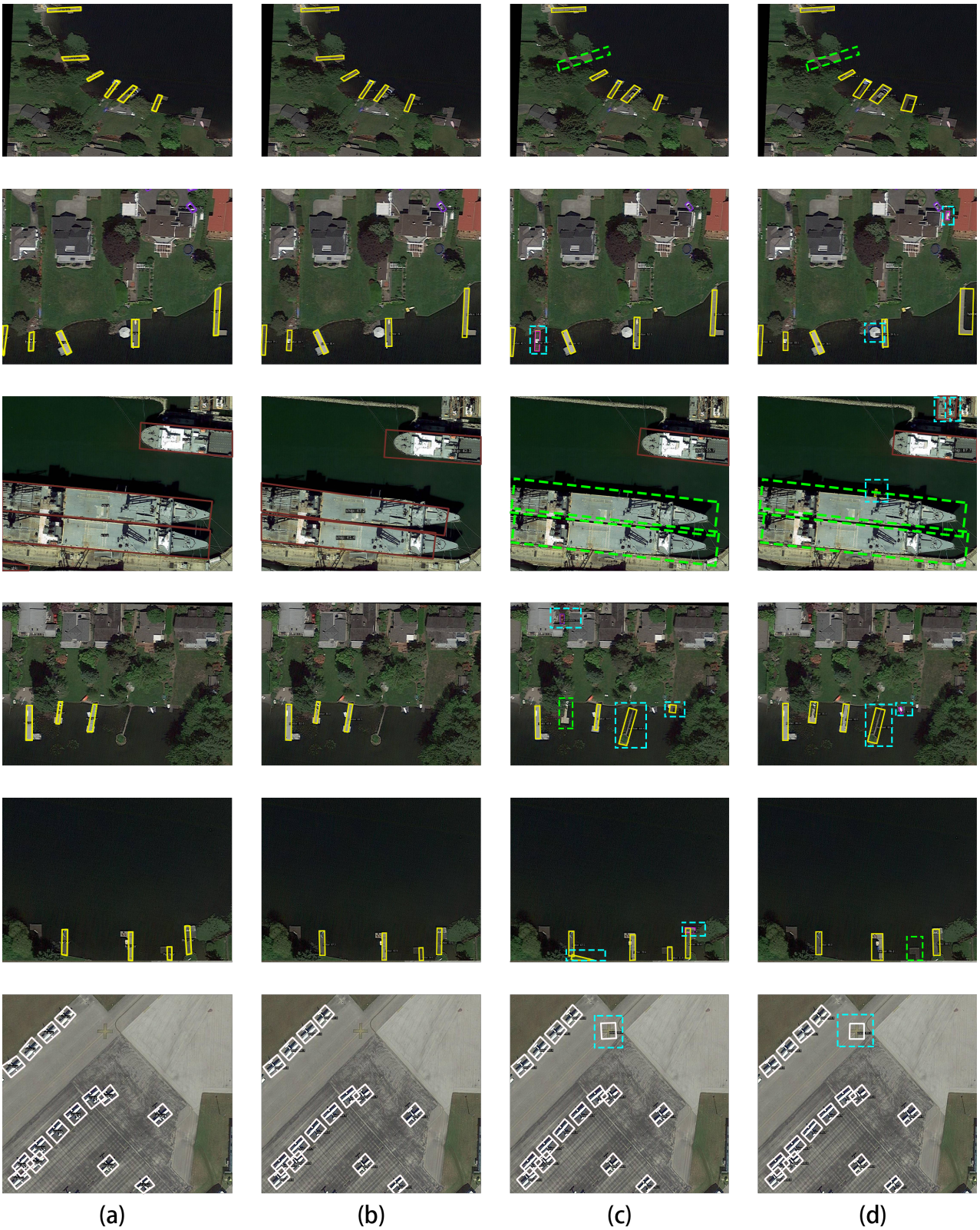

Figure 7. Results from various methods on the DOTA v1.0 dataset. (**a**) Ground truth. (**b**) MBSDet. (**c**) RTMDet. (**d**) ReDet. The dashed-green box indicates missed detection, and the dashed-cyan box indicates false detection. Solid rectangular boxes represent real or predicted results, where yellow indicates a harbor, white indicates a plane, red indicates a ship, and purple indicates a small vehicle.

Overall, MBSDet outperformed the HRSC2016 and DOTA v1.0 datasets in terms of object detection accuracy. Tables 1 and 2 show two top-performing comparison approaches for qualitative comparison with MBSDet. Qualitative and quantitative comparisons demonstrate that MBSDet can successfully tackle the issue of significant interference caused by complex backgrounds in aerial photographs.

3.4. Ablation

Ablation experiments on several modules with the HRSC2016 dataset were performed to illustrate the impact of our enhancements to RTMDet. In the RTMDet baseline representation, √ denotes the addition of the relevant module, while × indicates no module addition. Table 3 indicates that incorporating the MFE module during the feature fusion upsampling step captured additional details and localization information of objects. The MFE module increased detection accuracy by 1.2% mAP and 1.41% AP50. The BS module improved detection accuracy by 3.41% mAP and 4.84% AP50, demonstrating its ability to suppress background characteristics and improve resilience when detecting multi-scale targets.

Table 3. Findings from ablation experiments with different components.

MFE	BS	mAP	AP50
×	×	85.25	85.16
√	×	86.45	86.57
×	√	88.66	89.94
√	√	90.54	90.60

4. Discussion

Current methodologies have attempted to alleviate the influence of intricate backdrops on target detection. Spatial attention techniques have been utilized to improve target focus; yet, they struggle to be successful across multi-scale targets. This approach employs a Transformer-based mechanism for suppressing background features, focusing on highlighting regions of interest while reducing background noise. Some methods use super-resolution networks to preprocess images, which improves the representation of small multi-scale features; however, these networks are often complex and resource-demanding. This study presents the MFE (multi-focus enhancement) module to produce feature maps that are supplemented with additional contextual information. Comparative experimental results indicate that the proposed MBSDet attains enhanced performance.

4.1. BS Module

In the BS module, in addition to the suggested Transformer module, various other attention techniques were introduced. The SE attention module [48] increases the network's representational capabilities by explicitly modeling interdependencies among feature channels. The efficient multi-scale attention (EMA) module [49] reorganizes partial channels into the batch dimension to enhance feature representation. ACmix [50] integrates the global perception powers of self-attention mechanisms with the local feature extraction strengths of convolutions to augment the model's ability to process features. The four attention mechanisms were evaluated in experiments, as illustrated in Table 4. The experimental findings demonstrate that incorporating the Transformer attention (TA) mechanism in the BS module produces optimal performance.

To evaluate the efficacy of our BS module in mitigating the interference of background features, we performed a comparative analysis of the feature maps produced by the proposed multi-scale background suppression detection (MBSDet) model and the baseline model at the P3, P4, and P5 stages, as illustrated in Figure 8. Upon comparing the feature maps at these three stages, we noted that the baseline model is highly affected by background information, however, our model markedly concentrates attention on the target

items. Our model precisely represents the dimensions and form ratios of targets through heatmaps, irrespective of their size. The purpose of this paper is to design an end-to-end object detector that can effectively suppress features from complex backgrounds. Therefore, the analysis results indicate that TA is the best choice, even though it is not optimal in terms of computational load and complexity, which will be a problem we need to optimize in the future. The result described above demonstrates that our background suppression module effectively mitigates background features.

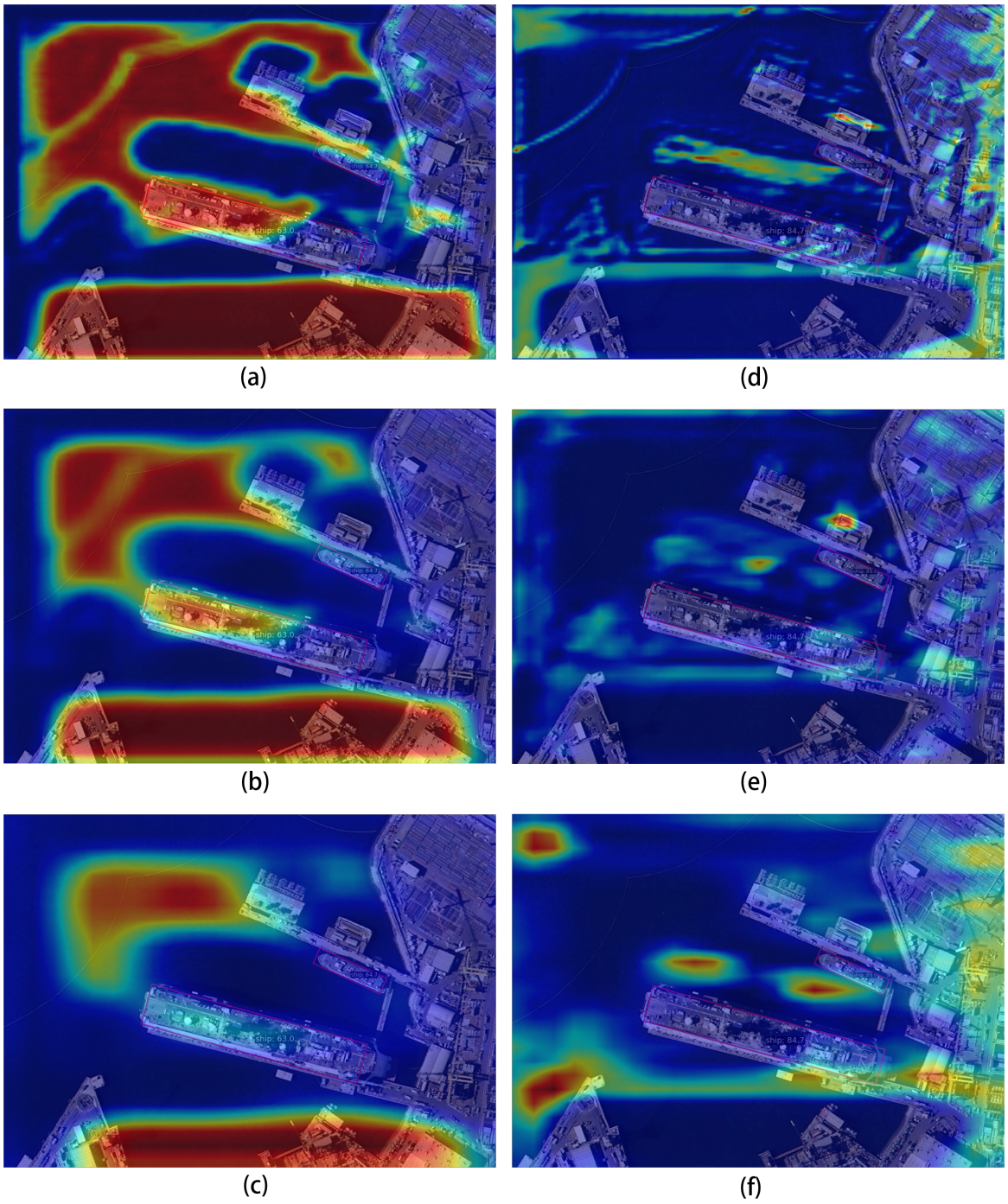

Figure 8. Feature map visualization results of the baseline and our MBSDet model. (**a**–**c**) represent the feature maps of the baseline model in P3, P4, and P5, respectively, and (**d**–**f**) represent the results of the corresponding phases of the MBS model.

Table 4. Analysis of several attention mechanisms within the BS module. The optimal outcomes are indicated in green.

Attention Model	mAP	AP50	Gflops	Params (M)
None	85.25	85.16	20.45	4.87
SE	82.41	83.40	20.45	4.92
EMA	85.54	86.50	20.98	4.9
ACmix	83.46	83.50	22.12	5.51
TA	90.54	90.60	22.42	6.51

4.2. Future Work

In future work, we would like to improve MBSDet in three principal domains. The incorporation of the BS module has augmented the number of parameters, leading to diminished computational speed; consequently, we will investigate the implementation of a more efficient Transformer model. Secondly, although the present work employs the DOTA v1.0 dataset, encompassing 15 categories including planes, ships, and storage tanks, we intend to isolate the ship category to develop a specialized subset and evaluate the model's detection ability on this subset. Finally, the HRSC2016 dataset from Northwestern Polytechnical University included in this work consists of high-resolution aerial photos that do not account for motion-blurred images. What we want to do is compile and simulate a novel dataset of blurred objects, and we will assess MBSDet's efficacy in identifying moving items utilizing this new dataset. Furthermore, we will explore innovative approaches to improve MBSDet's detection accuracy in aerial imagery with resolutions ranging from low to medium.

5. Conclusions

To reduce the impact of background features on target features in aerial images, we introduce an MBSDet detection model based on complex background feature suppression. The suggested BS module is intended to improve the multi-scale receptive field of features, which strengthens multi-scale target features while lowering extended background characteristics in complicated scenarios. The MFE module is specifically designed to achieve a more comprehensive feature representation by effectively extracting foreground features using feature pyramids. Extensive experimental results indicate the efficiency of MBSDet on publicly available datasets, for example, HRSC2016 and DOTA v1.0. Our MBSDet outperforms other leading models in terms of detection performance. We hope this work will provide useful insights to the community and inspire further study on exploring the potential of vision Transformers in additional computer vision tasks. We anticipate that our study will provide valuable insights for the community and stimulate further research into more efficient and precise methodologies for aerial image object recognition.

Author Contributions: G.J. wrote the Python source code, analyzed the data, and wrote the paper. C.L. analyzed the data and wrote the paper. L.F. helped with the project. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Intelligent Aerospace System Leading Innovation Team Program of Zhejiang (grant no. 2022R01003).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: We gratefully acknowledge the generous support provided by the Intelligent Aerospace System Leading Innovation Team.

Conflicts of Interest: All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Octavian, A.; Jatmiko, W. Designing intelligent coastal surveillance based on big maritime data. In Proceedings of the 2020 International Workshop on Big Data and Information Security (IWBIS), Depok, Indonesia, 17–18 October 2020; pp. 1–8.
2. Wu, J.; Li, R.; Li, J.; Zou, M.; Huang, Z. Cooperative unmanned surface vehicles and unmanned aerial vehicles platform as a tool for coastal monitoring activities. *Ocean Coast. Manag.* **2023**, *232*, 106421. [CrossRef]
3. Giusti, E.; Capria, A.; Saverino, A.L.; Gelli, S.; Mu noz-Casta ner, J.; Dosil, R.; Naya, J.; Menéndez, J. A Drone-Based Multisensory Payload for Maritime Pollutants Detections. *IEEE Aerosp. Electron. Syst. Mag.* **2022**, *38*, 4–18. [CrossRef]
4. Gonçalves, L.; Damas, B. Automatic detection of rescue targets in maritime search and rescue missions using uavs. In Proceedings of the 2022 International Conference on Unmanned Aircraft Systems (ICUAS), Dubrovnik, Croatia, 21–24 June 2022; pp. 1638–1643.
5. Zhang, Y.; Yin, Y.; Shao, Z. An Enhanced Target Detection Algorithm for Maritime Search and Rescue Based on Aerial Images. *Remote Sens.* **2023**, *15*, 4818. [CrossRef]
6. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
7. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
8. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14: Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
9. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
10. Xiang, A.; Zhang, L.; Fan, L. Shadow removal of spacecraft images with multi-illumination angles image fusion. *Aerosp. Sci. Technol.* **2023**, *140*, 108453. [CrossRef]
11. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
12. 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*, 5625411. [CrossRef]
13. Zhang, Z.; Peng, H.; Fu, J.; Li, B.; Hu, W. Ocean: Object-aware anchor-free tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 771–787.
14. Duan, K.; Xie, L.; Qi, H.; Bai, S.; Huang, Q.; Tian, Q. Corner proposal network for anchor-free, two-stage object detection. In Proceedings of the European Conference on Computer Vision. Springer, Glasgow, UK, 23–28 August 2020; pp. 399–416.
15. Lin, T. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
16. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
17. Han, J.; Ding, J.; Xue, N.; Xia, G.S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2786–2795.
18. Liu, J.; Qiao, W.; Xiong, Z. OAB-YOLOv5: One-Anchor-Based YOLOv5 for Rotated Object Detection in Remote Sensing Images. *J. Sens.* **2022**, *2022*, 8515510. [CrossRef]
19. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, SciTePress, Porto, Portugal, 24–26 February 2017; Volume 2, pp. 324–331.
20. Tian, Z.; Chu, X.; Wang, X.; Wei, X.; Shen, C. Fully convolutional one-stage 3d object detection on lidar range images. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 34899–34911.
21. Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; Yan, J. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18381–18394.
22. Yu, Y.; Da, F. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13354–13363.
23. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 3163–3171.
24. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
25. Xiang, A.; Zhang, L.; Fan, L. Design and analysis of an electro-adhesive hexapod robot with convertible limbs in microgravity. *Adv. Space Res.* **2024**, *73*, 1908–1924. [CrossRef]
26. Xiang, A.; Fan, L. Robust Spacecraft Image Interest Point Detection and Description with Transformer Under Variable Illumination Conditions. In Proceedings of the 2023 China Automation Congress (CAC), Chongqing, China, 17–19 November 2023; pp. 1235–1240.
27. 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.
28. Haixin, L.; Ziqiang, Z.; Zeyu, M.; Yeung, S.K. Marinedet: Towards open-marine object detection. *arXiv* **2023**, arXiv:2310.01931.
29. Bhalla, S.; Kumar, A.; Kushwaha, R. Analysis of recent techniques in marine object detection: A review. In *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–99.

30. Genin, L.; Champagnat, F.; Le Besnerais, G.; Coret, L. Point object detection using a NL-means type filter. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 3533–3536.
31. Kim, S.; Lee, J. Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track. *Pattern Recognit.* **2012**, *45*, 393–406. [CrossRef]
32. Kushwaha, A.K.S.; Srivastava, R. Framework for dynamic background modeling and shadow suppression for moving object segmentation in complex wavelet domain. *J. Electron. Imaging* **2015**, *24*, 051005. [CrossRef]
33. Kushwaha, A.K.S.; Srivastava, R. A framework for moving object segmentation using dynamic background modelling and shadow suppression in complex wavelet domain. *Imaging Sci. J.* **2016**, *64*, 267–278. [CrossRef]
34. Zheng, J.; Ren, J. Road Disease Detection based on Latent Domain Background Feature Separation and Suppression. *arXiv* **2023**, arXiv:2309.07616.
35. Pan, M.; Xia, W.; Yu, H.; Hu, X.; Cai, W.; Shi, J. Vehicle Detection in UAV Images via Background Suppression Pyramid Network and Multi-Scale Task Adaptive Decoupled Head. *Remote Sens.* **2023**, *15*, 5698. [CrossRef]
36. Huang, J.; Zhang, H.; Tang, G.; Bao, W. Robust UKF-based filtering for tracking a maneuvering hypersonic glide vehicle. *Proc. Inst. Mech. Eng. Part G: J. Aerosp. Eng.* **2022**, *236*, 2162–2178. [CrossRef]
37. Huang, J.; Li, Z.; Liu, D.; Yang, Q.; Zhu, J. An adaptive state estimation for tracking hypersonic glide targets with model uncertainties. *Aerosp. Sci. Technol.* **2023**, *136*, 108235. [CrossRef]
38. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef]
39. Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; Sutskever, I. Generative pretraining from pixels. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1691–1703.
40. Dosovitskiy, A. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
42. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
43. Zhang, Z.; Lu, X.; Cao, G.; Yang, Y.; Jiao, L.; Liu, F. ViT-YOLO: Transformer-based YOLO for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2799–2808.
44. Li, Y.; Wu, C.Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; Feichtenhofer, C. Mvitv2: Improved multiscale vision transformers for classification and detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4804–4814.
45. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
46. Liu, W.; Lu, H.; Fu, H.; Cao, Z. Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6027–6037.
47. Yu, Y.; Yang, X.; Li, Q.; Zhou, Y.; Zhang, G.; Yan, J.; Da, F. H2RBox-v2: Boosting hbox-supervised oriented object detection via symmetric learning. *arXiv* **2023**, arXiv:2304.04403.
48. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–13 June 2018; pp. 7132–7141.
49. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
50. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 815–825.

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