drones

MDPI

Article

Multi-Branch Parallel Networks for Object Detection in High-Resolution UAV Remote Sensing Images

Qihong Wu, Bin Zhang *, Chang Guo and Lei Wang

Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430205, China; 22107010012@stu.wit.edu.cn (Q.W.); 22107010006@stu.wit.edu.cn (C.G.); 22107010064@stu.wit.edu.cn (L.W.)
* Correspondence: zhangbin@wit.edu.cn

Abstract: Uncrewed Aerial Vehicles (UAVs) are instrumental in advancing the field of remote sensing. Nevertheless, the complexity of the background and the dense distribution of objects both present considerable challenges for object detection in UAV remote sensing images. This paper proposes a Multi-Branch Parallel Network (MBPN) based on the ViTDet (Visual Transformer for Object Detection) model, which aims to improve object detection accuracy in UAV remote sensing images. Initially, the discriminative ability of the input feature map of the Feature Pyramid Network (FPN) is improved by incorporating the Receptive Field Enhancement (RFE) and Convolutional Self-Attention (CSA) modules. Subsequently, to mitigate the loss of semantic information, the sampling process of the FPN is replaced by Multi-Branch Upsampling (MBUS) and Multi-Branch Downsampling (MBDS) modules. Lastly, a Feature-Concatenating Fusion (FCF) module is employed to merge feature maps of varying levels, thereby addressing the issue of semantic misalignment. This paper evaluates the performance of the proposed model on both a custom UAV-captured WCH dataset and the publicly available NWPU VHR10 dataset. The experimental results demonstrate that the proposed model achieves an increase in AP_L of 2.4% and 0.7% on the WCH and NWPU VHR10 datasets, respectively, compared to the baseline model ViTDet-B.

Keywords: UAV remote sensing images; object detection; self-attention; sampling; feature fusion

Citation: Wu, Q.; Zhang, B.; Guo, C.; Wang, L. Multi-Branch Parallel Networks for Object Detection in High-Resolution UAV Remote Sensing Images. *Drones* **2023**, *7*, 439. https://doi.org/10.3390/drones7070439

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 17 May 2023
Revised: 28 June 2023
Accepted: 30 June 2023
Published: 2 July 2023

1. Introduction

The evolution of Uncrewed Aerial Vehicle (UAV) technology has facilitated the acquisition of high-resolution remote sensing images. Object detection within UAV remote sensing images holds promise for applications in various sectors including urban planning, land monitoring, precision agriculture, updates to geographic information systems, and military operations, among others [1]. Nevertheless, the complexity of the background and the variability in size and orientation of objects in UAV remote sensing images introduce substantial challenges to object detection in UAV remote sensing images [2].

The advent of Convolutional Neural Networks (CNN) has significantly advanced the field of object detection. Current mainstream detection methodologies can be classified into two categories: one-stage and two-stage detection methods. The one-stage YOLO (You Only Look Once)-series algorithm prioritizes speed, directly regressing and classifying the object box of feature map predictions across various scales to enhance inference speed [3,4]. On the other hand, the two-stage RCNN-series algorithm excels in accuracy, utilizing the Region Proposal Network (RPN) to generate candidate boxes, which are then classified and regressed [5]. Inspired by these approaches, numerous researchers have integrated deep learning techniques into UAV remote sensing image object detection. To address the challenge of small-scale object detection, several novel methods have been proposed. Zeng et al. [6] proposed the SCA-YOLO (Spatial and Coordinate Attention enhancement YOLO) algorithm. A hybrid attention module with associated coordinate attention was designed by the algorithm to enhance the feature extraction of small objects. Furthermore, improvements were made to the SEB (Simple and Efficient Bottleneck) to further

distinguish the foreground and background characteristics. Zhou et al. [7] developed the ADCSPDarknet53 backbone network based on YOLOV4, modifying the regression loss function to enhance the model's ability to locate small objects. Zhuang et al. [8] proposed a one-stage detection model with multi-scale feature fusion, integrating different levels of feature maps for improved detection accuracy with small objects. Lan et al. [9] introduced a novel method for detecting small objects in optical remote sensing images. The proposed method constructs a spatial transformer that incorporates spatial attention and self-attention mechanisms. This facilitates the extraction of key features from relevant areas within the image space and mitigates the issue of small object leakage. Liu et al. [10] proposed the YOLO-extract algorithm, which is based on YOLOv5. The algorithm integrates coordinated attention into the network by adopting the concept of residual networks. Moreover, by combining the hybrid expansion convolution with the redesigned residual structure, the algorithm optimizes the model's feature extraction power for objects at various scales. Despite these advancements, CNN-based methods appear to have reached a performance plateau in detection tasks [11].

In recent years, numerous researchers have explored the representation of images. Considering images as a collection of patches has opened new avenues for research in object detection. Dosovitakiy et al. [12] proposed that transformers, possessing a capability for global-context information modeling and spatial adaptive aggregation, surpass the constraints of CNN and are widely employed in computer vision tasks [13–15]. The ViTDet [13] model utilizes the native ViT model as its backbone network for object detection. Notably, the ViTDet model is divided into four stages. The first few blocks of each stage use the window self-attention to improve computational efficiency, and the last block uses global self-attention to facilitate information exchange between different windows. This design renders the ViTDet model suitable for object detection tasks. Following this approach, Zhu et al. [16] proposed TPH-YOLOv5, a network that replaces the original prediction head with Transformer Prediction Heads (TPH). Simultaneously, a prediction head is added to detect objects at varying scales, thereby augmenting the model's object recognition capabilities. Zhang et al. [17] introduced the Transformers for Remote Sensing (TRS) model, incorporating self-attention into the ResNet network to improve the model's capacity to learn the overall features of the image and attain superior detection accuracy. Jiang et al. [18] presented the RAST-YOLO (You Only Look Once with Region Attention and Swin Transformer) algorithm. The algorithm uses the Region Attention (RA) mechanism combined with a Swin Transformer as the backbone to extract features to enhance the detection accuracy of objects in complex backgrounds. Subsequently, the C3D module is employed to fuse deep and shallow semantic information, thereby enhancing the detection accuracy of small targets. Wang et al. [19] introduced the MashFormer model, which combines a CNN with a transformer to enhance its representational ability in complex-background scenarios. Additionally, a multi-level feature aggregation module with cross-level feature alignment is designed to mitigate the semantic discrepancy between features extracted from shallow and deep layers.

The aforementioned method enhances the model's capability to extract object features through the use of a transformer, thereby effectively improving the detection performance. However, due to the advancements in UAV technology, high-resolution UAV remote sensing images have become easily obtainable. High-resolution UAV remote sensing images result in small objects that occupy fewer pixels within the image and are surrounded by complex background information. The network faces difficulty in extracting effective features from small objects, thereby impeding the model's ability to accurately locate and recognize them [20]. Furthermore, the process of multiscale prediction using an FPN (Feature Pyramid Network) [21] encounters the challenge of missing feature information. These challenges leave room for improving the detection performance of UAV remote sensing images.

To address the aforementioned issues, this paper employs the transformer structure to enhance the model's feature-encoding capability and mitigate the loss of semantic

information through the integration of diverse features. Specifically, this paper presents a Multi-Branch Parallel Network (MBPN) based on the ViTDet model. Initially, the object features from different feature maps input to FPN undergo enhancement through Receptive Field Enhancement (RFE) and Convolutional Self-Attention (CSA) modules. The RFE module is well-suited for shallow feature maps, enabling the extraction of features of varying sizes through convolutions with diverse kernel sizes. The resulting feature maps are concatenated along the channel dimension, while the original feature maps undergo convolutional and Softmax operations to generate attention maps. The CSA module is well-suited for deep feature maps, as it maps the feature maps to three distinct linear spaces (Q, K, V) through diverse convolution operations. Specifically, similarity calculations are performed by Q and K to derive the attention map, followed by weighting V to yield the final result. Additionally, the utilization of Multi-Branch Upsampling (MBUS) and Multi-Branch Downsampling (MBDS) modules yields diverse feature maps, which are then concatenated along the channel dimension and ultimately compressed through 1×1 convolution. Finally, the Feature-Concatenating Fusion (FCF) module is employed to merge the feature maps. This process involves sampling small-scale feature maps and concatenating them with large-scale feature maps, which are subsequently compressed through convolutional operations to yield the fused feature maps.

In summary, this paper contributes in the following ways:

(1) The introduction of the RFE and CSA modules into FPN enhances shallow and deep features, respectively, aiming to highlight the foreground and suppress noise interference.
(2) The MBUS and MBDS modules acquire diverse features through multiple paths, reducing the loss of feature information during the sampling process of the FPN.
(3) The FCF module fuses small-scale and large-scale feature maps, enriches feature information representation, and augments the semantic information of feature maps.

2. Related Work

The FPN employs a top-down structure with lateral connections to construct high-level semantic feature maps across multiple scales, thereby enhancing the flexibility of multi-scale representation and enjoying wide application in various detectors. However, it still presents certain limitations. For instance, there are semantic differences between layers, and direct fusion may diminish the power of multi-scale representation. Furthermore, feature information might be lost during the FPN network's sampling process. In this section, two key aspects will be explored: enhancing the ability of multi-scale representation and minimizing feature information loss.

Enhancing the Ability of Multi-scale Representation. Addressing the issue of the FPN difficulty in adapting to changes in object scale. Tang et al. [22] introduced the Scale-Aware Feature Pyramid Network (SARFNet). This approach employs 3-D convolution to establish a scale equilibrium pyramid convolution, enhancing the correlation between different feature levels and allowing flexible matching with objects exhibiting varying appearance changes. Additionally, Zhao et al. [23] proposed a multi-scale feature fusion module, named BFPCAR, which mitigates the imbalance of attention in non-adjacent layers of the FPN network. Dong et al. [24] innovatively replaced the lateral connection of the FPN with a deformable convolution lateral connection module to facilitate multi-scale object detection. Furthermore, Sun et al. [25] developed a Multi-Scale Feature Pyramid Network (MS-FPN) that amplifies shallow and deep features through the Atrophy Convolution Pyramid (ACP) module, while adaptively learning and selecting crucial feature maps using multi-scale attention modules.

Minimizing Feature Information Loss. To mitigate the loss of information during feature sampling and fusion, Chen et al. [26] introduced a Parallel Residual Dual Fusion Pyramid Network (PRB-FPN), designed to gather more comprehensive contextual information through bidirectional fusion. Furthermore, Guo et al. [27] applied the Residual Feature Augmentation module to counteract the loss of semantic information resulting from

channel reduction. This issue is addressed in this study through channel concatenation. Content-Aware Feature Reorganization (CARAFE) [28] generates multiple features in each feature map through various groups of content perception methods. Feature upsampling is then achieved by rearranging the generated features into a spatial block. This paper obtains multiple features through a variety of methods, concatenates the resulting features, and finally derives the final feature map using convolution operations. Zheng et al. [29] proposed the Gating Path Aggregation (GPA) network, asserting that different feature layers have varying degrees of importance. This network enhances the capability of information filtering during feature fusion.

3. Method

Figure 1 illustrates the structure of the object detection model presented in this study. Initially, the input image undergoes division into multiple image patches using the Patch operation. These image patches are subsequently fed into ViT with a block count of 12, resulting in the output feature layers L1, L2, L3, and L4 corresponding to blocks 3, 6, 9, and 12, respectively. In the next step, the shallow feature maps L1 and L2, as well as the deep feature maps L3 and L4, are separately fed into the RFE module and the CSA module. Thereafter, the application of MBUS and FCF modules yields four output feature layers, namely, {P1–P4}. Additionally, the P4 feature layer undergoes input into the MBDS module to generate a deeper feature map, denoted as P5, which aids in predicting larger objects. Afterwards, the feature map {P1–P5} is fed into the RPN network to generate the candidate box, and ultimately both the candidate box and feature map {P1–P5} are passed to the ROIHead module to obtain the classification and localization results.

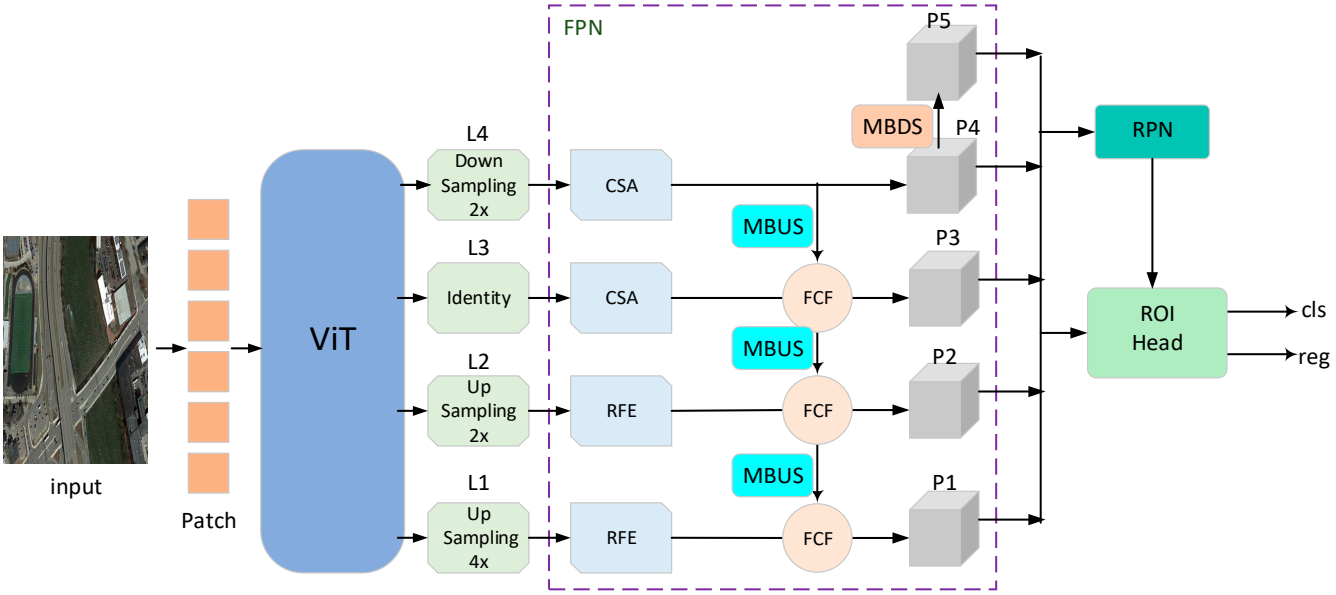

Figure 1. The structure of the proposed model.

3.1. RFE and CSA

UAV remote sensing images often contain noise originating from complex scenes, which can interfere with detection outcomes [2]. Additionally, the large feature map resolution of shallow networks tends to produce a lower level of feature abstraction and weaker semantic information, thereby containing more fine-grained details. On the other hand, deep networks further condense image information, enhancing the abstraction and semantics of features to better capture the image's overall characteristics. Hence, distinct foreground enhancement modules must be designed for different feature layers to suppress the impact of background information. Inspired by [30], this paper broadens the feature map's receptive field through convolutions of varying sizes, acquires object features as

weights, and subsequently applies these weights to the shallow feature map. Furthermore, this paper incorporate residual concepts [31] to bolster the model's generalization capacity. Deep features, characterized by high abstraction, necessitate the consideration of global information. Drawing inspiration from [12,32], this paper amplifies effective features by computing self-attention. The following sections provide a more detailed description of these two modules.

RFE. The specific design of the module is depicted in Figure 2. Given an input feature map, $X \in R^{C\times H\times W}$, this paper first acquire features $X1 \in R^{C\times H\times W}$, $X2 \in R^{C\times H\times W}$, and $X3 \in R^{C\times H\times W}$—representing small, medium, and large objects—through 3×3, 5×5, and 7×7 convolutions, respectively. Subsequently, $X1$, $X2$, and $X3$ are concatenated along the channel dimension. By applying a 1×1 convolution to reduce the channel count from 3C to C and using the Softmax function to generate the attention map, $attn \in R^{C\times H\times W}$. Following this, point-wise multiplication of $attn \in R^{C\times H\times W}$ and $X \in R^{C\times H\times W}$ is performed to derive the enhanced feature map, $Y \in R^{C\times H\times W}$. Finally, a residual branch is introduced to add $Y \in R^{C\times H\times W}$ and $X \in R^{C\times H\times W}$ to yield the final feature map, $Z \in R^{C\times H\times W}$. This process strengthens the attention of shallow networks towards multi-scale objects. The associated formula for the aforementioned procedure is presented below:

$$\begin{aligned} X1,\ X2,\ X3 &= Conv3(X),\ Conv5(X),\ Conv7(X) \\ attn &= Softmax(Conv1(Cat([X1,\ X2,\ X3]))) \\ Z &= attn\ *\ X\ + X \end{aligned} \tag{1}$$

where Conv3, Conv5, Conv7 are convolution operations with kernel size 3, 5, 7; Cat operation is concatenating in the channel dimension; and $*$ represents point multiplication.

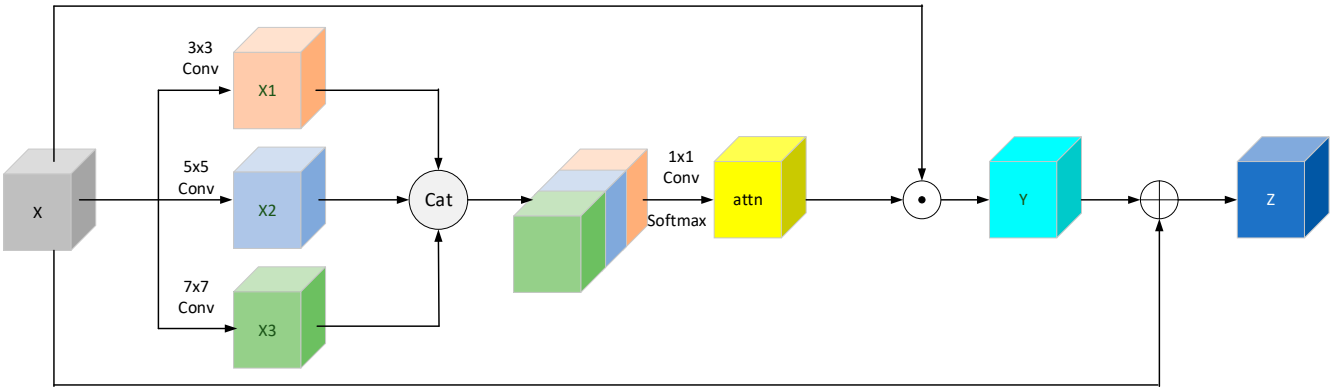

Figure 2. The structure of the RFE module.

CSA. The deep feature enhancement module is shown in Figure 3. Initially, the input feature map $X \in R^{C\times H\times W}$ undergoes a convolution with three 3×3 kernels, resulting in three spatial feature maps: $Q \in R^{C\times H\times W}$, $K \in R^{C\times H\times W}$, and $V \in R^{C\times H\times W}$. In the subsequent step, $Q \in R^{C\times H\times W}$ and $K \in R^{C\times H\times W}$ are adjusted to generate new feature maps $Q \in R^{L\times C}$ and $K \in R^{C\times L}(L = H \times W)$, respectively. Following this, Q and K perform matrix multiplication and the Softmax function is applied to yield the attention map $attn \in R^{L\times L}$. In the third phase, the dimension of $V \in R^{C\times H\times W}$ is adjusted to produce a new feature map $V \in R^{C\times L}(L = H \times W)$, which then undergoes matrix multiplication with 'attn' in conjunction with V to generate $Y \in R^{C\times L}$. Lastly, $Y \in R^{C\times L}$ employs a linear mapping layer, adjusts its dimension, and merges it with the residual branch $X \in R^{C\times H\times W}$ to generate the final feature map $Z \in R^{C\times H\times W}$. The formula representing the above procedure is presented below:

$$\begin{aligned} Q,\ K,\ V &= \ Conv3(X),\ Conv3(X),\ Conv3(X) \\ Q,\ K &= \ Q.reshape(\ C,\ L).permute\ (1,\ 0),\ K.reshape(\ C,\ L) \\ attn &= Softmax(Q\ @\ K) \\ V &= \ V.reshape(\ C,\ L) \\ Z &= \ (Linear(V\ @\ attn)).reshape(C,\ H,\ W)\ +\ X \end{aligned} \tag{2}$$

where Conv3 represents convolution operations with kernel size 3×3, and the three convolutions in the formula do not share parameters; @ is matrix multiplication; and Linear is a linear mapping layer.

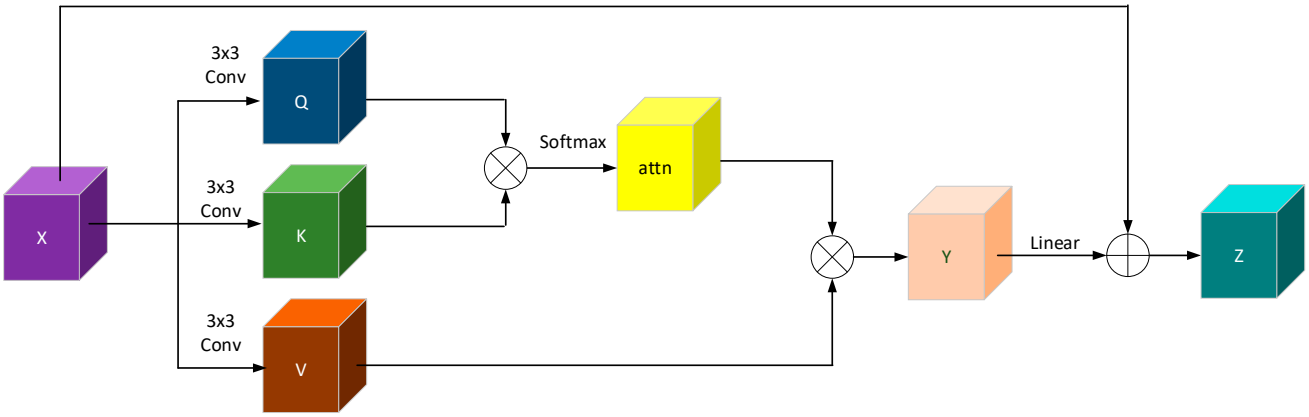

Figure 3. The structure of the CSA module.

3.2. MBUS and MBDS

In the domain of remote sensing image object detection, multi-scale features enhance the model's detection accuracy [2]. However, some semantic information is lost during the sampling operation performed for feature fusion. Drawing inspiration from [33], this paper designed multi-branching paths in this study to extract multiple features from feature map, thereby mitigating the loss of semantic information during the upsampling process. Provided below is a comprehensive description of the upsampling and downsampling modules devised in this paper.

MBUS. The most prevalent upsampling methods include nearest interpolation, bilinear interpolation, and transposed convolution. The upsampling kernel is determined by the spatial position of the pixels in the nearest or bilinear upsampling. The two regions of interest measure 1×1 and 2×2, respectively [28], and characteristics of objects of varying sizes are captured through these different-sized regions. Additionally, features are also extracted via learnable transposed convolution upsampling. This study considers these three methods of characteristic extraction to alleviate the issue of information loss. The detailed design of this method is depicted in Figure 4. For the input feature map $X \in R^{C \times H \times W}$, three feature maps ($X1 \in R^{C \times 2H \times 2W}$, $X2 \in R^{C \times 2H \times 2W}$, and $X3 \in R^{C \times 2H \times 2W}$) are generated via transposed convolution, nearest interpolation, and bilinear interpolation. These are then concatenated in the channel dimension, and finally, through a 1×1 convolution, the upsampled feature map $Y \in R^{C \times 2H \times 2W}$ is obtained. The formula to express this operation is presented below:

$$\begin{aligned} X1,\ X2,\ X3 &= \ ConvT(X),\ Nearest(X),\ Bilinear(X) \\ Y &= Conv1(Cat([X1,\ X2, X3])) \end{aligned} \tag{3}$$

where ConvT is transposition convolution, Nearest is nearest interpolation, Bilinear is bilinear interpolation, Cat is concatenated in the channel dimension, and Conv1 is 1×1 convolution operation.

MBDS. The design principles for downsampling mirror those of upsampling. This paper incorporates three downsampling methods, namely, convolution downsampling, maximum pooling, and average pooling. These methods generate multiple features for the newly created feature maps. The intricate design of this procedure is illustrated in Figure 5. For the input feature map $X \in R^{C\times 2H\times 2W}$, three distinct feature maps ($X1 \in R^{C\times H\times W}$, $X2 \in R^{C\times H\times W}$, and $X3 \in R^{C\times H\times W}$) are produced through the convolution operation with a kernel size of 3 and a stride of 2, maximum pooling, and average pooling. Subsequently, these maps are concatenated in the channel dimension, and through a 1x1 convolution, the downsampled feature map $Y \in R^{C\times H\times W}$ is obtained. The formula to express this operation is presented below:

$$\begin{aligned} X1,\ X2,\ X3 &= \ Conv3\_2(X),\ MaxPooling(X),\ AvgPooling(X) \\ Y &= Conv1(Cat([X1,\ X2, X3])) \end{aligned} \tag{4}$$

where Conv3_2 is a convolution with kernel 3 and step size 2, Cat is stitched in the channel dimension, and Conv1 is a convolution with 1×1.

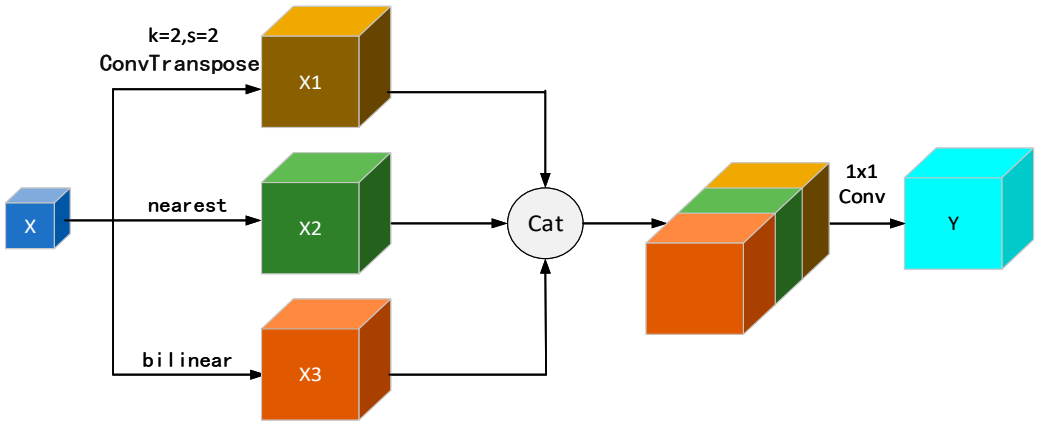

Figure 4. The structure of the MBUS module.

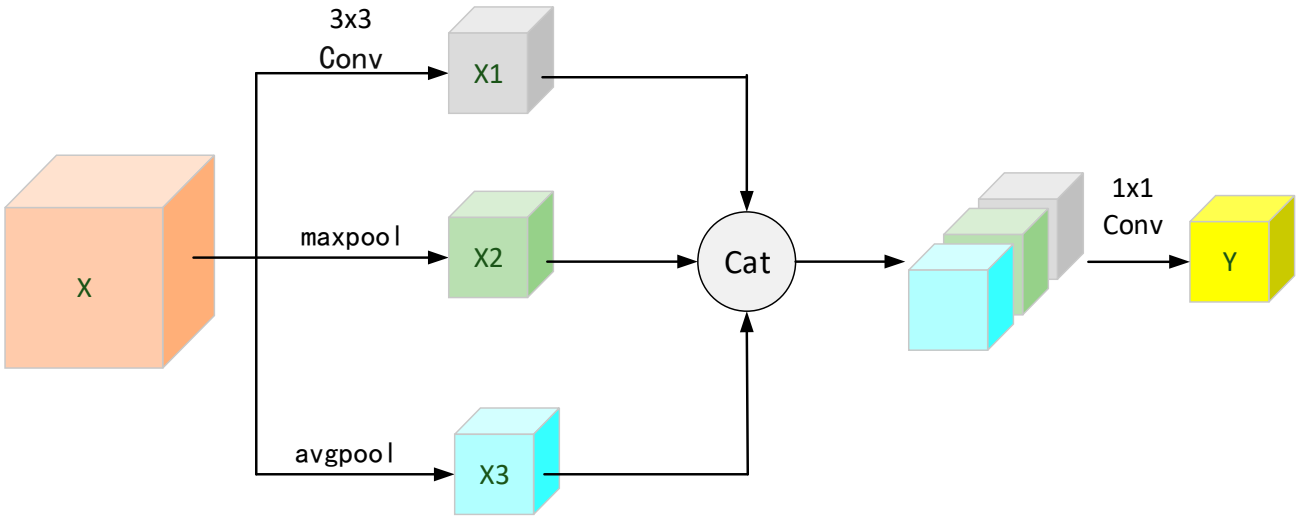

Figure 5. The structure of the MBDS module.

3.3. FCF

Common fusion methods encompass addition and concatenation. While the dimension of the feature map remains unchanged with additive fusion, the information within the feature map increases. On the other hand, concatenative fusion expands the dimension

to describe more features. During multiscale fusion, the information varies across different layers, making concatenative fusion a more reasonable choice than additive fusion. The detailed design of this module is illustrated in Figure 6. Here, the two input feature maps are concatenated in the channel dimension, followed by the interaction of the information from the two feature maps via a 3×3 convolution. The process concludes with a dimension reduction using a 1×1 convolution. The corresponding formula is expressed as follows:

$$Y = Conv1(Conv3(Cat([X1, X2]))) \quad (5)$$

where Cat is a concatenate operation, Conv3 is the convolution of kernel 3, and Conv1 is the convolution of kernel 1.

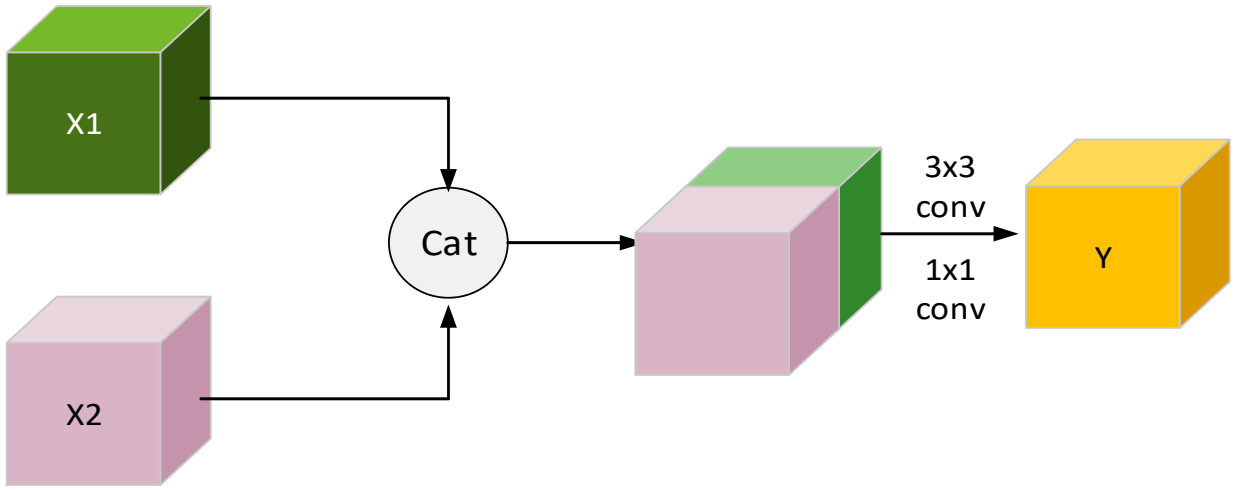

Figure 6. The structure of the FCF module.

4. Experiment

This section provides an overview of the datasets used in this study, the applied evaluation metrics, and experimental details. Subsequently, ablation experiments were performed on each module developed in this work to determine the contribution of each module to the performance enhancement. Finally, to validate the detection performance of the proposed model, this paper compares it with multiple methods on WCH and NWPU VHR10 datasets.

4.1. Datasets and Evaluation Metrics

In this paper, experiments are conducted on our own dataset WCH and the publicly available dataset NWPU VHR10 [34–36], the details of which are as follows:

WCH. This dataset's images are derived from aerial drone photography of Caidian District, Wuhan, suitable for UAV remote sensing image object detection. Due to the high resolution of the captured images, this paper cropped them to generate 1344 new 640×640 resolution images. After annotating the cropped images, a total of 32,349 instances covering one category are obtained, with each image containing multiple instances of arbitrary size and orientation. This paper randomly divided this dataset in an 8:2 ratio, resulting in 1075 images for training and 269 images for validation.

NWPU VHR10. This dataset's images are sourced from Google Earth and the Vaihingen dataset, which consists of aerial drone photography from Vaihingen, Germany. The latter is a subset of the test data used by the German Association of Photogrammetry and Remote Sensing (DGPF) for digital aerial cameras. The NWPU VHR-10 dataset, annotated using Horizontal Bounding Boxes (HBB), is publicly accessible and suitable for object detection in UAV remote sensing images. This paper omits unlabeled images from the NWPU VHR-10 dataset, retaining 650 images and 3896 instances across ten categories. The

images range from 400 × 500 to 1100 × 1800 in size. The dataset, characterized by variable object sizes and orientations, presents significant challenges. Given that the NWPU VHR10 dataset does not segregate training and validation sets, this paper calculated the image count for each category, dividing the images in an 8:2 ratio, resulting in 521 training images and 129 validation images.

Moreover, this paper used the AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L metrics to evaluate the detection performance of the model. AP represents the average precision across 10 intersection over union (IoU) thresholds ranging from 0.5 to 0.95, with intervals of 0.05. AP_{50} denotes the average precision at an IoU threshold of 0.5. AP_{75} represents the average precision at an IoU threshold of 0.75. AP_S indicates the average precision for small objects. AP_M signifies the average precision for medium objects. AP_L represents the average precision for large objects. Among these metrics, AP corresponds to the area under the precision–recall (P-R) curve, where P stands for precision and R stands for recall, as defined by the following formula:

$$\begin{aligned} P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{TP+FN} \\ AP &= \int_0^1 P(R)dR \end{aligned} \tag{6}$$

where TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively; P(R) is the precision–recall curve.

4.2. Implementation Details

This study's experiments employ ViTDet, which is implemented based on the MMDetection framework, as the baseline model. The proposed model used the pretrained weights on the ImageNet [37] dataset and initialized the remaining model parameters randomly. During training, the input image size was adjusted to 704 × 704 as part of data preprocessing, followed by random image cropping and flipping. The batch size is set to 2, the initial learning rate is 0.0001, and a linear warm-up strategy is used for the first 500 iterations. The model was trained for 30 epochs with a learning rate decay by a factor of 10 at the 15th and 25th epochs. The AdamW optimizer was used with beta coefficients set at (0.9, 0.999) and a weight decay of 0.1. All experiments were executed on an Ubuntu 20.0 system, with training accelerated by an NVIDIA GeForce RTX 4080 graphics card.

4.3. Ablation Experiments

Ablation experiments were conducted on the WCH dataset to assess the effectiveness of the proposed modules in this paper. To ensure a fair comparison, the hyperparameters for all ablation experiments were set according to the specifications outlined in Section 4.2. Subsequently, the RFE, CSA, MBUS, MBDS, and FCF modules were individually added to the baseline model (ViTDet-B) for experimentation. The results of the ablation experiments are presented in Table 1. In Table 1, "√" indicates that the module was added, while "×" indicates its absence. The first row displays the results of experiments conducted on the baseline model. In this table, the red font signifies a decrease in the indicator, while the green font signifies an increase in the indicator.

Table 1. Ablation experiments for all design modules in this paper on the WCH dataset.

RFE	CSA	MBUS	MBDS	FCF	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	×	×	×	×	**60.0**	86.8	68.1	**37.5**	**68.2**	64.8
√	×	×	×	×	59.6 (−0.4)	86.9 (+0.1)	67.7 (−0.4)	36.8 (−0.7)	68.0 (−0.2)	65.2 (+0.4)
×	√	×	×	×	59.5 (−0.5)	**87.0** (+0.2)	67.7 (−0.4)	36.9 (−0.6)	67.6 (−0.6)	63.6 (−1.2)
×	×	√	×	×	59.8 (−0.2)	**87.0** (+0.2)	**68.4** (+0.3)	37.0 (−0.5)	68.0 (−0.2)	**66.1** (+1.3)
×	×	×	√	×	59.1 (−0.9)	86.8 (+0.0)	67.2 (−0.9)	36.5 (−1.0)	67.4 (−0.8)	64.5 (−0.3)
×	×	×	×	√	59.6 (−0.4)	**87.0** (+0.2)	68.2 (+0.1)	37.1 (−0.4)	67.7 (−0.5)	65.6 +(0.8)

Ablation on RFE module. The RFE module is applied to the shallow feature layers L1 and L2. Convolution operations with varying kernel sizes are performed to obtain feature maps with multi-scale information, enabling adaptation to multi-scale objects in UAV remote sensing images. The results in the second row of Table 1 demonstrate that the utilization of the RFE module leads to an increase of 0.1% and 0.4% in AP_{50} and AP_L, respectively. This indicates that the RFE module significantly enhances the detection accuracy of large objects. However, the AP_S and AP_M of this module still fall behind those of the baseline model. This could be attributed to the disturbance caused to the features of small and medium objects in the shallow layer when the deep features are fused additively with the shallow features.

Ablation on CSA module. The module is applied to the deep feature layers L3 and L4 to enhance feature expression through self-attention calculation and spatial position weighting of the feature map. The utilization of the CSA module results in a 0.2% increase in AP_{50}, as observed in the third row of Table 1. This indicates that the model achieves improved accuracy in classifying and locating certain objects. Nevertheless, other indicators remained below the baseline level. This could be attributed to CSA enhancing both object features and noise features, particularly in complex scenes where the background occupies a significant portion of the area in UAV remote sensing images. Subsequently, a top-down fusion path is employed, which extends the distribution range of noise features, resulting in an unsatisfactory detection effect of the model.

Ablation on MBUS module. The purpose of this module is to upsample a smaller-sized feature map into a larger-sized feature map. Specifically, multiple feature-extraction branches are employed to acquire diverse feature information from high-level feature maps, which are subsequently utilized to construct high-level feature maps. The fourth row of Table 1 reveals that the utilization of the MBUS module leads to a 0.3% and 1.3% increase in AP_{50} and AP_L, respectively. These improvements can be attributed to the combination of diverse abstract features and positioning information. However, there was a slight decrease in AP_S and AP_M. This is because the MBUS module introduces additional background noise to the shallow feature layer, thereby impeding the model's localization ability and degrading its performance.

Ablation on MBDS module. The purpose of this module is to replace the original pooling operation and mitigate the loss of semantic information during pooling. The findings from the fifth row of Table 1 indicate that incorporating additional semantic information into the construction of the P5 feature map does not enhance the model's performance. This could be attributed to the introduction of significant background noise into the P5 feature map by MBDS, thereby resulting in an unsatisfactory detection effect of the model. In contrast, the baseline model employs the pooling operation to generate the P5 feature map. While this approach results in the loss of certain semantic information, it also discards some noise information, mitigating the impact of noise on the model.

Ablation on FCF module. This module concatenates two feature maps along the channel dimension and utilizes convolutional operations to facilitate the interaction between spatial and channel information. The results from the sixth row of Table 1 demonstrate that the inclusion of the FCF module in the baseline model leads to an improvement of 0.2%, 0.1%, and 0.8% in AP_{50}, AP_{75}, and AP_L, respectively. The effectiveness of the FCF module is confirmed. However, there was a slight decrease in AP_S and AP_M. This could be attributed to the FCF module covering the features of small and medium objects during the information exchange process, while the features of large objects are retained due to their larger spatial coverage.

4.4. Comparisons Experiments

Comparisons experiments were conducted on the WCH and NWPU VHR10 datasets to assess the performance of the proposed object detection model in this paper. This paper compare the proposed model with various object detection models, including one-stage mainstream models YOLOv7 and YOLOv8, two-stage classical models Faster RCNN

and Cascade RCNN, and representative transformer-series models Swin Transformer and ViTDet (the baseline model in this paper). The hyperparameters of the proposed model align with those described in Section 4.2, while the comparison models are implemented based on MMDetection and MMYOLO, respectively.

4.4.1. Comparison on the WCH Dataset

Table 2 presents the experimental outcomes obtained from the proposed model and the comparative model when applied to the WCH dataset. The proposed model demonstrates an increase of 0.1% and 2.4% in AP and AP_L, respectively, compared to the baseline model (ViTDet-B). These results indicate that the model proposed in this study enhances object positioning ability and exhibits improved perception of large objects. The ablation experiment results in Table 1 further confirm these observations, attributing them to the utilization of the RFE, MBUS, and FCF modules. The RFE module enhances object features, the MBUS module enables the acquisition of diverse features, and the FCF module effectively fuses these diverse features. Notably, the proposed model achieves an AP_{50} increase of 1.4% and 1.5% when compared to Faster RCNN and Cascade RCNN, respectively. This improvement can be attributed to the powerful encoding capabilities of the transformer. While the proposed model outperforms the one-stage object detection model—namely, YOLOv7—a significant gap remains between the proposed model and YOLOv8. Additionally, the proposed model surpasses Swin Transformer in terms of detection performance, specifically by improving the AP_L by 5%. This is potentially due to the fact that Swin Transformer employs local self-attention to reduce computational overhead and uses sliding windows for information propagation between different windows, whereas the proposed model employs global attention to propagate information, thereby surpassing the spatial information propagation limitations for enhanced performance.

Table 2. Comparison of detection performance in the WCH dataset. "*" denotes unpublished papers.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv7 [4]	57.9	84.3	65.8	34.7	66.9	52.1
YOLOv8 *	**65.3**	**87.5**	**74.1**	**41.8**	**74.1**	**71.6**
Faster RCNN [5]	55.8	85.2	63.9	35.0	63.6	60.2
Cascade RCNN [38]	60.5	85.1	68.7	38.1	68.6	61.7
Swin Transformer [14]	59.3	87.4	67.4	37.2	67.4	62.2
ViTDet-B [13]	60.0	86.8	68.1	37.5	68.2	64.8
Proposed	60.1	86.6	68.2	37.2	68.2	67.2

Figure 7 presents the visualization of the detection results achieved by the proposed model and the comparative model on the WCH dataset. The first column presents scenes characterized by a sparse background and a dense distribution of objects. The second column showcases scenes with a more prominent background. The third column displays scenes with objects of varying colors, and the fourth column depicts scenes where the background and objects exhibit similarities. Each row corresponds to the detection results of a specific model. The presence of a red circle indicates a missed detection object, whereas yellow circles indicate objects that the proposed model successfully detects but other models fail to identify. Additionally, the prediction results have been obtained with a confidence level set to 0.8. Figure 7 reveals that all models encounter the issue of missed detections. Notably, the YOLOv8 model outperforms other models in quantitative analysis, yet its visual detection results are unsatisfactory. This observation can be attributed to the low confidence level of YOLOv8 predictions, which is understandable considering its emphasis on detection speed. Moreover, the proposed model demonstrates stronger competitiveness compared to other models, particularly in scenarios involving occluded objects. This finding highlights the ability of the proposed model to enhance feature perception.

4.4.2. Comparison on the NWPU VHR10 Dataset

The experimental results of the model proposed in this paper and the comparison model on the NWPU VHR10 dataset are presented in Table 3. The model in this paper demonstrates an improvement over the baseline model ViTDet-B, with an increase of 1.8% and 0.7% in AP_{75} and AP_L, respectively, indicating enhanced detection performance for large objects. This finding aligns with the experimental results in Tables 1 and 2. However, there has been a decline in other indicators, particularly a 1.0% decrease in AP_S. The ablation experiment reveals that this decline can be attributed to the limited perception ability of the improved module in this paper when detecting small objects in UAV remote sensing images. In comparison to YOLOv7, YOLOv8, Faster RCNN, and Cascade RCNN, the model proposed in this paper exhibits slight advantages in AP_{50}, AP_{75}, and AP_L. In contrast to Swin Transformer, the model in this paper achieves a slightly lower AP_{50} score, which could be attributed to the advantage of Swin Transformer's local self-attention mechanism.

Table 3. Comparison of detection performance in the NWPU VHR10 dataset. "*" denotes unpublished papers.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv7 [4]	55.4	89.0	64.7	**55.1**	53.7	58.1
YOLOv8 *	62.4	95.3	70.6	48.5	58.0	71.0
Faster RCNN [5]	55.4	93.4	59.1	33.9	51.0	61.3
Cascade RCNN [38]	65.3	93.9	78.7	51.5	60.6	69.5
Swin Transformer [14]	**65.9**	**97.1**	77.1	45.1	**62.0**	71.8
ViTDet-B [13]	64.9	95.7	78.0	51.5	60.0	71.9
Proposed	64.6	95.5	**79.8**	50.2	59.5	**72.6**

Figure 8 presents the visualized detection results of the model proposed in this paper and the comparison model on the NWPU VHR10 dataset. The figure consists of four columns: the first column depicts scenes with a large object distribution, the second column portrays scenes with complex backgrounds and dense objects, the third column illustrates scenarios with small object distribution, and the fourth column represents scenes with large object distribution and redundant background information. Each row corresponds to the detection results of a specific model. The presence of a red circle denotes a missed detection object, while a yellow rectangle signifies that the model proposed in this paper detects it, while most other models do not. Furthermore, the prediction results are evaluated with a confidence level set to 0.8. Figure 8 reveals that the models from the transformer series exhibit a lower rate of missed objects. The third column of Figure 8 demonstrates that, in comparison to the baseline model ViTDet-B, the proposed model performs better in detecting objects in shadowed scenes, but its performance is suboptimal in scenes where the object closely resembles the background. This discrepancy may arise from the superior capability of the RFE and CSA modules in distinguishing objects from dissimilar backgrounds, while struggling to differentiate backgrounds that closely resemble objects.

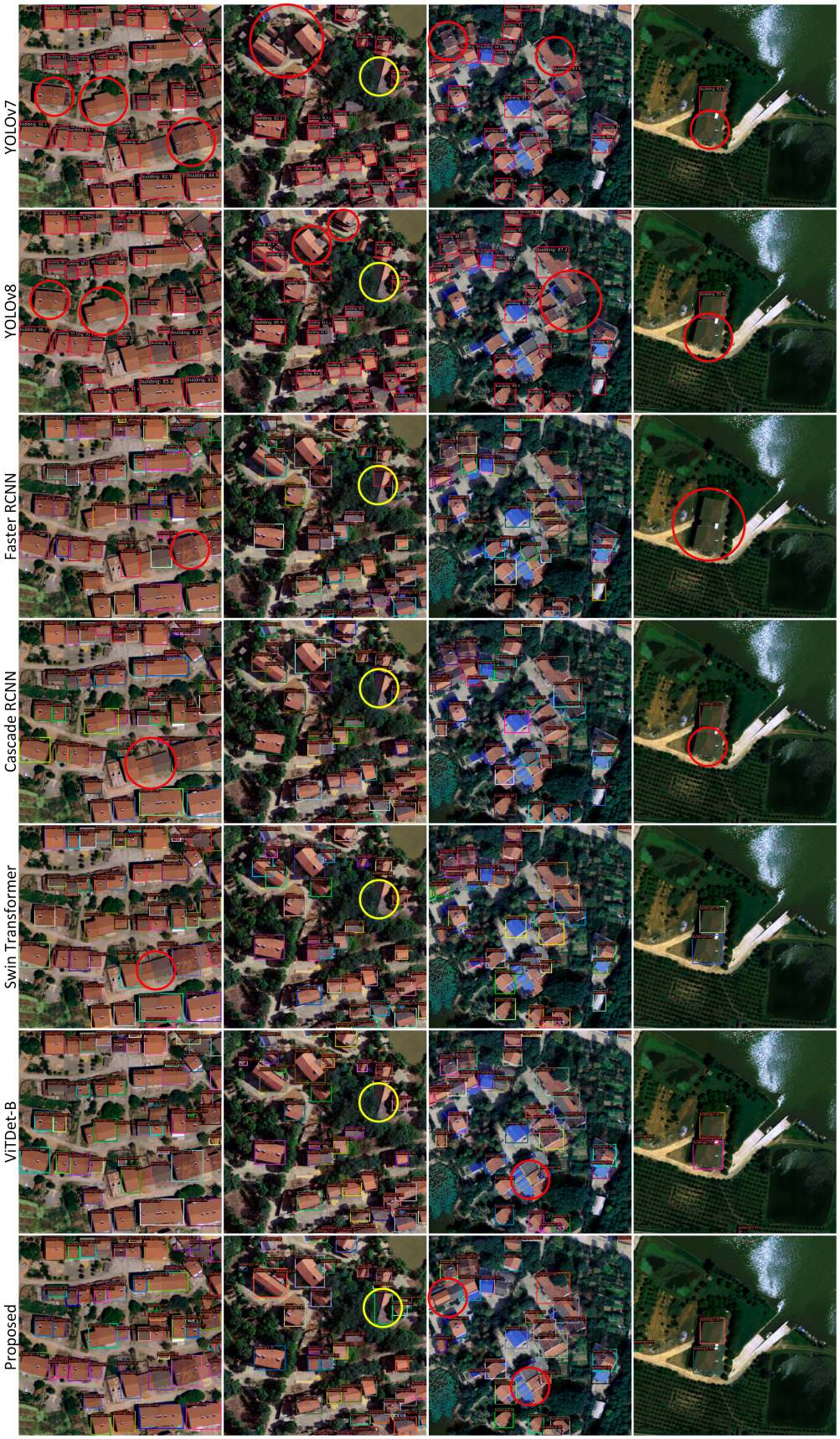

Figure 7. Visualization of the detection results on the WCH dataset.

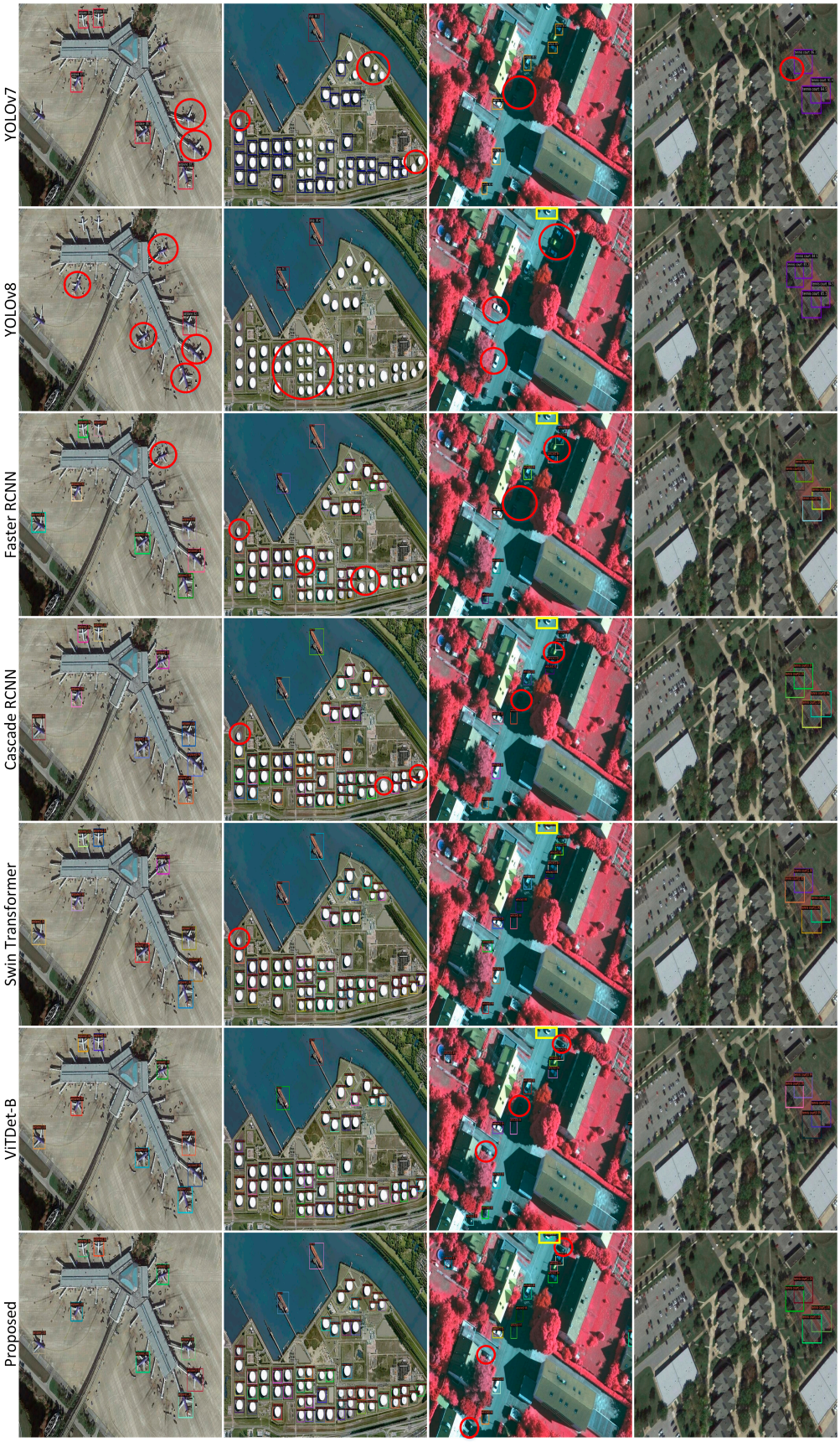

Figure 8. Visualization of the detection results on the NWPU VHR10 dataset.

5. Discussion

Ablation experiments on the WCH dataset are conducted to examine the influence of the modules proposed in this paper on the detection performance of the model. The results of the ablation experiments indicate that the designed module exhibits improvements in AP_{50} but leads to declines in other indicators. This could be attributed to the deliberate design of each module to focus on improving specific indicators rather than multiple indicators simultaneously. The comparative experiments in Section 4.4 demonstrate that the model presented in this paper outperforms the comparison model in detecting large objects. Furthermore, the proposed model exhibits impressive detection performance in scenes featuring occluded objects (Second column in Figure 7) and shadowed scenes (Third column in Figure 8). This can be attributed to the RFE module's successful expansion of the object's receptive field in the shallow feature map and the CSA module's enhancement of the weight of the object feature.

Nevertheless, the model presented in this paper is suboptimal for detecting small and medium objects. This could be due to the presence of noise information in the L1, L2, L3, and L4 feature layers generated by the backbone network. While the RFE module is capable of filtering out certain noise from the shallow features through convolution, the CSA module inadvertently amplifies the eigenvalue of the noise when assigning weights to the object features in the deep feature map. Consequently, during the top-down fusion process, a portion of the noise from the deep feature maps is reintroduced into the shallow feature maps, thereby impacting the model's detection performance.

Based on the above observations, in the task of detecting objects in UAV remote sensing images, it is imperative to progressively reduce the noise information within the deep feature map as the network becomes deeper, thereby enhancing the model's detection performance.

6. Conclusions

The presence of complex background information and densely distributed objects in UAV remote sensing images can adversely affect the model's detection performance. To address this issue, the present paper introduces the MBPN model, which enhances the FPN by making improvements. Initially, the RFE and CSA modules enhance the feature representation of foreground objects. Subsequently, the MBUS and MBDS modules mitigate the loss of semantic information during FPN sampling. Lastly, the FCF module alleviates the problem of semantic information misalignment during the feature fusion process.

Ablation experiments validate the efficacy of the proposed module in this study. Furthermore, comparative experiments conducted on the WCH and NWPU VHR10 datasets demonstrate the high competitiveness of the proposed method. Nevertheless, the model presented in this paper still exhibits certain limitations. For instance, in the ablation experiment, the enhanced module displays improvements in some evaluation indicators while experiencing decreases in others. Additionally, in the comparative experiment, the model demonstrates suboptimal detection performance for small and medium objects.

Subsequent research will involve conducting more comprehensive investigations aimed at enhancing the model's detection accuracy for small objects.

Author Contributions: Q.W. was responsible for collecting and labeling image data, designing experiments, performing experiments, analyzing experiments, and writing the manuscript; B.Z. provided writing guidance; C.G. and L.W. were responsible for labeling image data. All the authors made important contributions to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Science Foundation of Hubei Province of China under the Grant 2022CFCO31; and the Discipline Innovation and Intelligence Introduction Program for Colleges and Universities under the Grant B17040.

Data Availability Statement: The datasets used in this paper are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johson, B. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
2. Dong, X.; Qin, Y.; Fu, R.; Gao, Y.; Liu, S.; Ye, Y.; Li, B. Multiscale deformable attention and multilevel features aggregation for remote sensing object detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
3. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
4. Wang, C.Y.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 18–22 June 2023; pp. 7464–7475. [CrossRef]
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2015**, *28*, 1137–1149. [CrossRef]
6. Zeng, S.; Yang, W.; Jiao, Y.; Geng, L.; Chen, X. SCA-YOLO: A new small object detection model for UAV images. *Vis. Comput.* **2023**, *39*, 1–17. [CrossRef]
7. Zhou, H.; Ma, A.; Niu, Y.; Ma, Z. Small-Object Detection for UAV-Based Images Using a Distance Metric Method. *Drones* **2022**, *6*, 308. [CrossRef]
8. Zhuang, S.; Wang, P.; Jiang, B.; Wang, G.; Wang, C. A single shot framework with multi-scale feature fusion for geospatial object detection. *Remote Sens.* **2019**, *11*, 594. [CrossRef]
9. Lan, J.; Zhang, C.; Lu, W.; Gu, N. Spatial-Transformer and Cross-Scale Fusion Network (STCS-Net) for Small Object Detection in Remote Sensing Images. *J. Indian Soc. Remote Sens.* **2023**, *51*, 1–13. [CrossRef]
10. Liu, Z.; Gao, Y.; Du, Q.; Chen, M.; Lv, W. YOLO-Extract: Improved YOLOv5 for Aircraft Object Detection in Remote Sensing Images. *IEEE Access* **2023**, *11*, 1742–1751. [CrossRef]
11. Li, Q.; Chen, Y.; Zeng, Y. Transformer with transfer CNN for remote-sensing-image object detection. *Remote Sens.* **2022**, *14*, 984. [CrossRef]
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010. [CrossRef]
13. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring plain vision transformer backbones for object detection. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 280–296. [CrossRef]
14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision--ECCV 2020: 16th European Conference (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229. [CrossRef]
16. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 10–17 October 2021; pp. 2778–2788. [CrossRef]
17. Zhang, J.; Zhao, H.; Li, J. TRS: Transformers for remote sensing scene classification. *Remote Sens.* **2021**, *13*, 4143. [CrossRef]
18. Jiang, X.; Wu, Y. Remote Sensing Object Detection Based on Convolution and Swin Transformer. *IEEE Access* **2023**, *11*, 38643–38656. [CrossRef]
19. Wang, K.; Bai, F.; Li, J.; Liu, Y.; Li, Y. MashFormer: A Novel Multiscale Aware Hybrid Detector for Remote Sensing Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2753–2763. [CrossRef]
20. Wang, J.; Shao, F.; He, X.; Lu, G. A Novel Method of Small Object Detection in UAV Remote Sensing Images Based on Feature Alignment of Candidate Regions. *Drones* **2022**, *6*, 292. [CrossRef]
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
22. Tang, L.; Tang, W.; Qu, X.; Han, Y.; Wang, W.; Zhao, B. A scale-aware pyramid network for multi-scale object detection in SAR images. *Remote Sens.* **2022**, *14*, 973. [CrossRef]
23. Zhao, Y.; Li, J.; Li, W.; Shan, P.; Wang, X.; Li, L.; Fu, Q. MS-IAF: Multi-Scale Information Augmentation Framework for Aircraft Detection. *Remote Sens.* **2022**, *14*, 3696. [CrossRef]
24. Dong, X.; Qin, Y.; Gao, Y.; Fu, R.; Liu, S.; Ye, Y. Attention-Based Multi-Level Feature Fusion for Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3735. [CrossRef]
25. Sun, Z.; Meng, C.; Cheng, J.; Zhang, Z.; Chang, S. A Multi-Scale Feature Pyramid Network for Detection and Instance Segmentation of Marine Ships in SAR Images. *Remote Sens.* **2022**, *14*, 6312. [CrossRef]

26. Chen, P.Y.; Chang, M.C.; Hsieh, J.W.; Chen, Y.S. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE Trans. Image Process.* **2021**, *30*, 9099–9111. [CrossRef] [PubMed]
27. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 12595–12604. [CrossRef]
28. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 3007–3016. [CrossRef]
29. Zheng, Y.; Zhang, X.; Zhang, R.; Wang, D. Gated Path Aggregation Feature Pyramid Network for Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 4614. [CrossRef]
30. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400. [CrossRef]
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
32. Yu, S.; Xiao, J.; Zhang, B.; Lim, E.G. Democracy does matter: Comprehensive feature mining for co-salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–23 June 2022; pp. 979–988. [CrossRef]
33. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017; p. 31. [CrossRef]
34. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
35. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
36. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
37. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
38. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef]