FFCA-YOLO for Small Object Detection in Remote Sensing Images

Yin Zhang, Mu Ye, Guiyi Zhu, Yong Liu, Pengyu Guo, and Junhua Yan

***Abstract*— Issues, such as insufficient feature representation and background confusion, make detection tasks for small object in remote sensing arduous. Particularly, when the algorithm will be deployed on board for real-time processing, which requires extensive optimization of accuracy and speed under limited computing resources. To tackle these problems, an efficient detector called feature enhancement, fusion and context aware YOLO (FFCA-YOLO) is proposed in this article. FFCA-YOLO includes three innovative lightweight and plug-and-play modules: feature enhancement module (FEM), feature fusion module (FFM), and spatial context aware module (SCAM). These three modules improve the network capabilities of local area awareness, multiscale feature fusion, and global association cross channels and space, respectively, while trying to avoid increasing complexity as possible. Thus, the weak feature representations of small objects are enhanced and the confusable backgrounds are suppressed. Two public remote sensing datasets (VEDAI and AI-TOD) for small object detection and one self-built dataset (USOD) are used to validate the effectiveness of FFCA-YOLO. The accuracy of FFCA-YOLO reaches 0.748, 0.617, and 0.909 (in terms of mAP50) that exceeds several benchmark models and the state-of-the-art methods. Meanwhile, the robustness of FFCA-YOLO is also validated under different simulated degradation conditions. Moreover, to further reduce computational resource consumption while ensuring efficiency, a lite version of FFCA-YOLO (L-FFCA-YOLO) is optimized by reconstructing the backbone and neck of FFCA-YOLO based on partial convolution (PConv). L-FFCA-YOLO has faster speed, smaller parameter scale, and lower computing power requirement but little accuracy loss compared with FFCA-YOLO. The source code will be available at https://github.com/yemu1138178251/FFCA-YOLO.**

***Index Terms*— Context information, feature fusion, lightweight network, remote sensing image, small object detection.**

I. Introduction

In RECENT years, the research on small object detection has achieved significant growth due to the rapid development of optical remote sensing technology [1], [2], [3], [4], [5], [6] for applications, such as traffic supervision, search and rescue, security, military, and so on. Remote sensing images generally have large fields of view, which is quite suitable for wide area monitoring. However, because of their relatively low resolution and poor quality, interested objects are usually characterized by small sizes (less than 32×32 pixels [7], [57]), dim features, low contrast, and insufficient information, causing extra difficulties in detection [8], [9]. At the same time, remote sensing systems face less controllable observing conditions and numerous interferences in imaging chain, such as platform motion, atmosphere, and various complex imaging scenes. All these factors lead to the aliasing of objects and backgrounds, which makes small objects indistinguishable. On the other hand, with the continuous increase of camera bands and resolution, massive data are generated during on-board imaging [10]. For example, WorldView-4 collect data covering 680 000 km^2 per day [11], which brings a huge amount of downstream data. Traditional ground processing mode after data downlink is facing severe challenges, which is hard to meet the requirements of high timeliness applications, such as military reconnaissance and emergency rescue. Real-time processing on board can significantly relieve transmission pressure of imaging data and shorten the delay from information acquisition to strategic decision, which becomes one of the potential ways to solve this problem. Authoritative institutions, such as European Space Agency (ESA), have already treated on-board processing technology as one of the key research directions prospectively [12]. Unfortunately, the strict constraints on on-board resources, such as power, weight, and volume, put forward higher requirements for the performance of processing algorithms in terms of reliability, speed, and scale.

In general, the main challenges of small object detection in remote sensing applications can be summarized into three points: insufficient feature representation, background confusion, and the optimization of speed and accuracy under limited hardware conditions.

In this study, our motivation is to design a small object detector with high accuracy that has the potential to be applied to real-time processing on board in the future. The key to alleviate the problems of insufficient feature representation and background confusion lies in feature enhancement and fusion. In terms of feature enhancement, fully utilizing local and global contextual information [13], [14], [15] can effectively enhance the perception of network for small objects. Feature enhancement module (FEM) and spatial context aware module (SCAM) are proposed to enrich the local and global contextual feature, respectively. FEM expands the receptive

Manuscript received 23 September 2023; revised 16 December 2023; accepted 31 January 2024. Date of publication 6 February 2024; date of current version 28 February 2024. This work was supported in part by the Strengthening Project of National Defense Science and Technology under Grant 2021-JCJQ-JJ-0834, in part by the Nation Natural Science Foundation of China under Grant 61705104, in part by the Fundamental Research Funds for the Central Universities of China under Grant NJ2022025. *(Corresponding authors: Yin Zhang; Pengyu Guo.)*

Yin Zhang, Mu Ye, Guiyi Zhu, and Junhua Yan are with the College of Astronautics, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing 211106, China (e-mail: zhangyin1986@nuaa.edu.cn; yemu_nuaa@nuaa.edu.cn; guiyi_zhu@nuaa.edu.cn; yjh9758@126.com).

Yong Liu and Pengyu Guo are with the National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing 100071, China (e-mail: xhliuyong@sina.com; pengyu.guo@nudt.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3363057

field of the backbone by multibranch atrous convolution. SCAM considers the association between small objects and global regions by constructing global context relationships. In terms of feature fusion, feature fusion module (FFM) is proposed to improve feature fusion strategy, which can reweight different feature maps by channel information without increasing computational complexity. These three modules are added to YOLO to obtain a new model: feature enhancement, fusion, and context aware YOLO (FFCA-YOLO). Finally, in order to further reduce computational resource consumption while ensuring efficiency, a lite version of FFCA-YOLO (L-FFCA-YOLO) is optimized by reconstructing the backbone and neck of FFCA-YOLO based on partial convolution (PConv).

The main contributions of this article are listed as follows.

1) An efficient detector (FFCA-YOLO) of small objects and its lite version L-FFCA-YOLO are designed for remote sensing applications. FFCA-YOLO has advanced performance in small object detection tasks compared with several benchmark models and the state-of-the-art (SOTA) methods, and has the potential for future real-time application on board.
2) Three innovative and lightweight plug-and-play modules are proposed: FEM, FFM, and SCAM. These three modules improve the network capabilities of local area awareness, multiscale feature fusion, and global association cross channels and space, respectively. They can be used as common modules inserting into any detection networks to enhance the weak feature representations of small objects and suppress the confusable backgrounds.
3) A new small object dataset USOD is constructed based on aerial remote sensing images, which has the proportion of small objects (less than 32×32 pixels) more than 99.9% with many instances under low illumination and shadow occlusion conditions. In addition, USOD has multiple test sets under different simulated degradation conditions, such as image blurring, Gaussian noise, stripe noise, and fog, which can serve as a benchmark dataset for small object detection in remote sensing.

The remainder of this article is organized as follows: after introducing the related works of small object detection in Section II, the proposed FFCA-YOLO and L-FFCA-YOLO architecture are elaborated in Section III. In Section IV, the experimental details are briefly introduced. The performance of the proposed method and several benchmark models as well as SOTA methods are particularly compared. The robustness and lightweight performance of FFCA-YOLO are also validated in this section. In Section V, the entire article is summarized and the future directions of small object detection in remote sensing are pointed out.

II. Related Works

This section briefly reviews the literatures relevant to our work, including the applications of YOLO in remote sensing detection, feature extraction methods of small object, global context feature representation, and lightweight frameworks of network.

A. Applications of YOLO in Remote Sensing

The development of deep learning enables object detectors to adaptively extract image features and locate objects through end-to-end learning framework. At present, the detection methods can be classified into two categories: two-stage [16], [17] and one-stage detectors [18], [19], [20], [21]. Compared with two-stage detectors, one-stage detectors have faster computation speed and low accuracy loss, which makes them have better potential for on-board applications. YOLO series of algorithm [18], [19], [20], as typical one-stage object detection algorithms, has advantages to achieve desired performance for small objects. At present, some improved YOLO algorithms for object detection in remote sensing have emerged, such as TPH-YOLO [22], FE-YOLO [23], and CA-YOLO [24].

TPH-YOLO [22] integrates transformer encoder blocks into backbone to obtain rich global context information and improves the quality of object feature representation. FE-YOLO [23] uses deformable convolution for feature fusion of high and low feature maps in the neck of YOLO, which aims to eliminate the impact of semantic gaps caused by top–down connections on objects. These two methods have good results but with a sharp increase in parameter count. CA-YOLO [24] embeds coordinate attention module into shallow feature network extraction, which suppresses redundant backgrounds and enhances the feature representation of objects by establishing long-range dependencies between pixels. In summary, YOLO has the superiority of scalability and efficiency, which is suitable for applying in remote sensing tasks.

Therefore, we choose YOLO as the basic framework and add specifically designed modules for small object feature representation and background suppression.

B. Feature Enhancement and Fusion Methods of Small Object Detection

Object detection methods based on deep learning rely on the backbone to obtain high-dimensional features. However, in remote sensing images, the extracted features of small objects may only occupy one pixel on output feature maps. Multiscale features need to be used to represent the features more effectively. Inspired by the pyramid structure derived from hand-engineered features, Lin et al. [25] propose the feature pyramid network (FPN), which yields the capacity to aggregate low-level features that have high resolution with high-level features that have low resolution. Since then, PANet [26], NAS-FPN [27], ASFF [28], and BiFPN [29] are proposed and achieve good results in object detection tasks. Guo et al. [30] introduce AugFPN to address the inconsistency between detailed and semantic information in feature maps. The information gap is narrowed by using a one-time supervision method in feature fusion stage. Liu et al. [31] present a high-resolution object detection network (HRDNet) to detect small vehicle objects, which uses a multidepth image pyramid combined with a multiscale FPN to deepen features. These methods demonstrate that strengthening the quality of multiscale feature fusion can effectively improve the detection performance of small objects to a certain extent. In addition,

feature enhancement before fusion can further improve the semantic representation of network. Cheng et al. [32] use dual attention mechanism to enhance features before fusion, which makes the network focus on the distinct features of objects. The feature enhance module proposed by Zhang and Shen [33] is similar to Cheng's, which also uses the attention mechanism of spatial and channel dimensions to enhance features. Besides attention mechanism, expanding the receptive field by multibranch convolution [8] and transformer encoder [34], [35] are also two commonly used ways for feature enhancement.

In order to obtain a larger receptive field, a new lightweight FEM is designed for obtaining richer local contextual information in this article, which includes a multibranch structure containing standard convolution and atrous convolution. In addition, a new FFM is proposed by improving the multiscale fusion strategy with almost no additional parameters.

C. Global Context Feature Representation

After FEM and FFM, the feature representation of small objects has been enhanced to some extent. Modeling the global relationship between small objects and backgrounds at this stage is more effective than in backbone.

According to the research results of [36], [37], and [38], obtaining the global receptive field and context information is very important for small object localization. Nonlocal neural network (NLNet) [13] aggregates the global context by calculating the pairwise correlations between spatial pixels. After that, GCNet [14] and SCP [38] simplify the multiplication of query and key to solve the problem of excessive calculation of NLNet. SCP adds additional paths to GCNet to learn the information of each pixel. This additional path uses one 1×1 convolution to aggregate spatial information between different channels, which may still bring some useless background features.

Based on these methods, a new SCAM is proposed considering the ideas of [39] and [40]. SCAM uses global average-pooling (GAP) and global max-pooling (GMP) to guide pixels learning the relationship between space and channels. Therefore, the proposed SCAM can achieve contextual feature interaction cross channels and space.

D. Lightweight Model Frameworks

Lightweight is an important indicator for measuring detector performance, especially aiming at on-board deployment in the future, which requires to optimize accuracy and speed with limited computing resources. There are two commonly used ways to make network lightweight. The first one is model compression represented by pruning [41], [42], [43], [44]. The essence of pruning is to delete the redundant parameters lower than the threshold set by designing filtering algorithm. Any model can be pruned to reduce the amount of parameters. Another way is to use lightweight convolutional networks to optimize the model structure. Its idea lies in designing more efficient computing methods for networks. MobileNet [45], ShuffleNet [46], and GhostNet [47] use the depthwise convolution (DWConv) and/or group convolution to extract spatial features. DWConv can effectively reduce parameter count and FLOPs. Several network structures [48], [49], [50] for object detection in remote sensing implement lightweight design based on the above methods. Chen et al. [51] prove that the low FLOPs of DWConv are mainly due to frequent memory access by operators. Therefore, the PConv is proposed to extract the spatial features more effectively by reducing redundant calculations and memory access. Based on the idea of PConv, a lite version of FFCA-YOLO named L-FFCA-YOLO is presented by reconstructing the network in Section IV-E, which is faster and slightly lower in accuracy.

III. Proposed Method

A. Overview

YOLOv5 is selected as our benchmark framework since it has fewer parameters compared with the latest YOLOv8 and can maintain a certain degree of accuracy in the tasks of small object detection. The overall architecture of FFCA-YOLO is shown in Fig. 1. First, FFCA-YOLO only uses four convolution subsampling operations as the backbone of feature extraction, which is different from the original YOLOv5. Second, three specially designed modules are added into the neck of YOLOv5: a lightweight FEM is proposed to improve the local area awareness of the network; FFM is proposed to improve the capability of multiscale feature fusion; SCAM is designed to improve the capability of global association cross channels and space. Finally, a lite version named L-FFCA-YOLO is obtained by reconstructing FFCA-YOLO based on PConv with little accuracy loss. Their detailed description can be found in Sections III-B–III-E.

B. Feature Enhancement Module (FEM)

Due to the complexity of remote sensing images, false alarms with similar features are prone to occur in tasks of small object detection. However, the extraction ability of backbone is limited. The features extracted at this stage contain less semantic information and narrow receptive fields, which makes it difficult to distinguish small objects from backgrounds. Accordingly, the proposed FEM considers to enhance the features of small objects from two perspectives. From the view of increasing feature richness, multibranch convolutional structure is adopted to extract multiple discriminative semantic information. From the view of enlarging receptive fields, atrous convolution is applied to obtain richer local contextual information. The whole structure of FEM is shown in Fig. 2, which is inspired by RFB-s [52]. The difference is that FEM only has two branches with atrous convolution. Each branch performs a 1×1 convolution operation on the input feature map to preliminarily adjust the number of channels for subsequent processing. The first branch is a residual structure, which forms an equivalent map to retain critical feature information of small objects. The other three branches perform cascade standard convolution operations, whose kernel sizes are 1×3, 3×1, and 3×3, respectively. Additional atrous convolution layers are added to the middle two branches, so that the extracted feature maps could retain more context information.

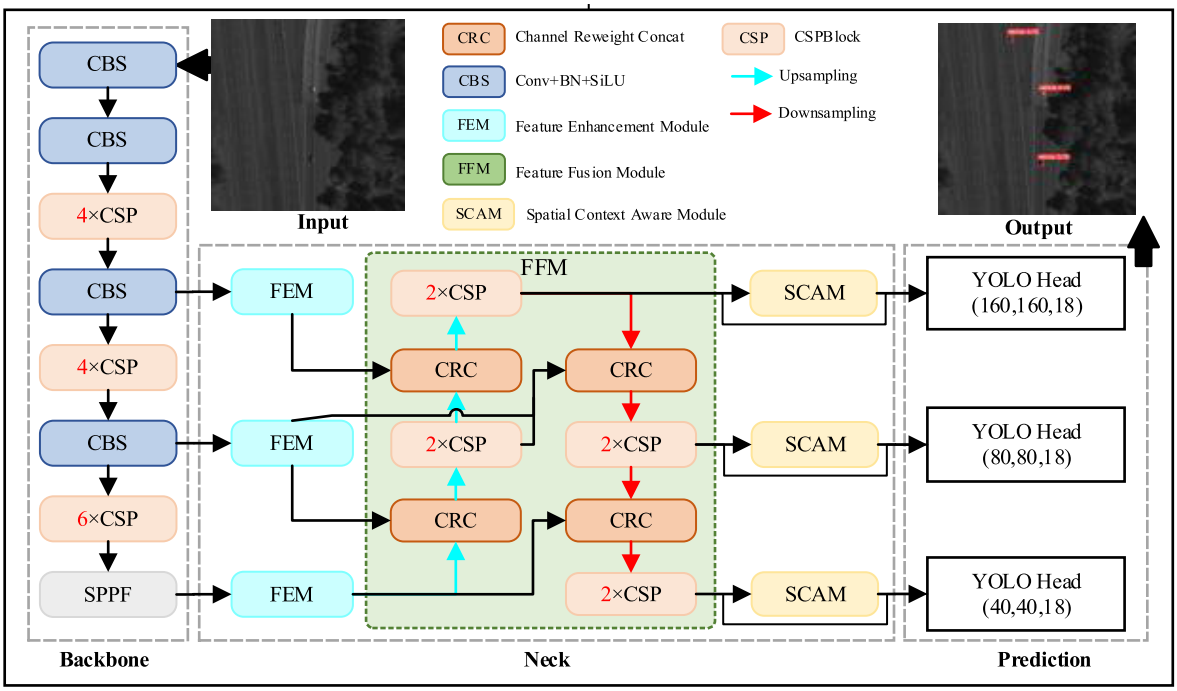


Fig. 1. Overall framework of FFCA-YOLO.

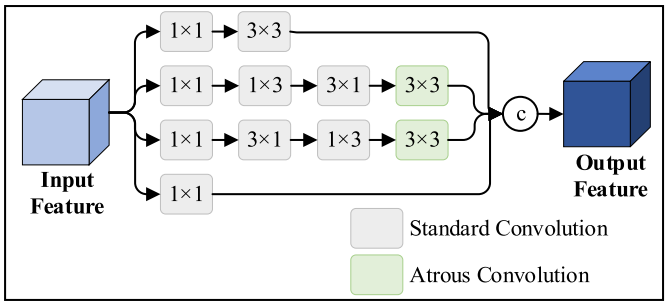


Fig. 2. Structure of FEM.

The mathematical expressions of FEM can be written as follows:

$$W_1 = f_{\text{conv}}^{3\times3}\left[f_{\text{conv}}^{1\times1}(F)\right] \tag{1}$$

$$W_2 = f_{\text{diconv}}^{3\times3}\left\{f_{\text{conv}}^{3\times1}\left\{f_{\text{conv}}^{1\times3}\left[f_{\text{conv}}^{1\times1}(F)\right]\right\}\right\} \tag{2}$$

$$W_3 = f_{\text{diconv}}^{3\times3}\left\{f_{\text{conv}}^{1\times3}\left\{f_{\text{conv}}^{3\times1}\left[f_{\text{conv}}^{1\times1}(F)\right]\right\}\right\} \tag{3}$$

$$Y = \text{Cat}(W_1, W_2, W_3) \oplus f_{\text{conv}}^{1\times1}(F) \tag{4}$$

where $f_{\text{conv}}^{1\times1}$, $f_{\text{conv}}^{1\times3}$, $f_{\text{conv}}^{3\times1}$, and $f_{\text{conv}}^{3\times3}$ represent the standard convolution operations with kernel sizes of 1×1, 1×3, 3×1, and 3×3, respectively. $f_{\text{diconv}}^{3\times3}$ means atrous convolution operation with a dilation rate of 5. Cat$(\cdot)$ is the feature map concatenation operation. $\oplus$ represents the elementwise addition operation of the feature map. F is the input feature map. W_1, W_2, and W_3 represent the output feature map of the first three branches after standard and atrous convolution. Y is the output feature map of FEM.

FEM has a much lighter structure compared with RFB-s and enables the model to learn richer local contextual features through multibranch atrous convolution, which improves the feature representation ability for small objects.

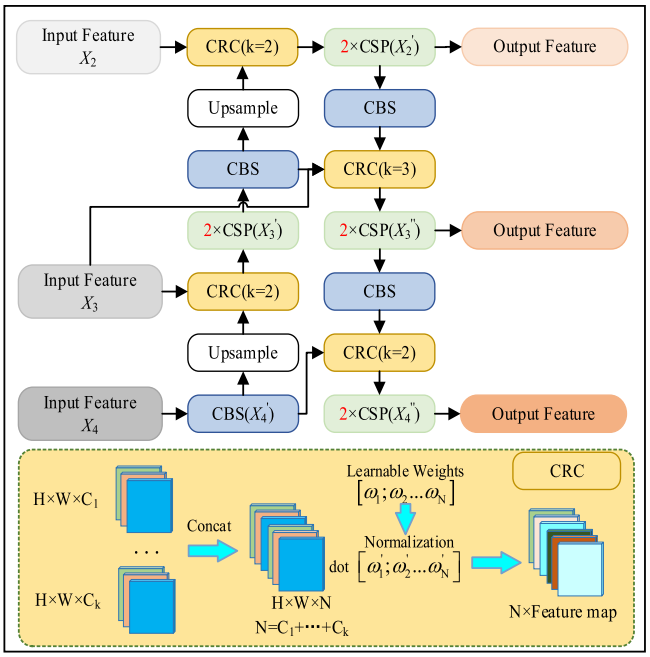


Fig. 3. Structure of FFM.

C. Feature Fusion Module (FFM)

High-level and low-level feature maps contain different semantic information. Aggregating features from multiscale feature maps could enhance the semantic representation of small object. The proposed FFM adopts a neck structure based on BiFPN. Unlike BiFPN, FFM improves the reweighting strategy named CRC and adjusts the original BiFPN to accommodate three detection heads. The structure of FFM is shown in Fig. 3. The input of FFM consists of the low-level feature

maps $X_2(160 \times 160)$ and $X_3(80 \times 80)$ processed by FEM and the high-level feature map $X_4(40 \times 40)$ processed by SPPF.

The top–down strategy of FFM is as follows. First, using CSPBlock for X_4 to get X'_4, then upsampling X'_4 to obtain the feature map with the same scale as X_3, and using CRC to fuse them together. The fused feature map is processed by CSPBlock to get X'_3. The above operations are repeated on X'_3 to create a new feature map X'_2. X'_2, X'_3, and X'_4 realize the flowing of semantic information from deep to shallow. The process from bottom to top is similar to that from top to bottom, with the main difference being that the feature map is downsampled using a convolution with a stride of 2. X''_3 is obtained through the CRC of X_3, X'_3, and X'_2. This operation could fuse more features without increasing much costs. X'_2, X''_3, and X''_4 as the output results of FFM are sent to SCAM for context information extraction. The calculation process of FFM can be expressed as follows:

$$X'_2 = \mathrm{CSP}\left\{\mathrm{CRC}\left[f_{\mathrm{up}}^{2\uparrow}\left(\mathrm{CBS}(X'_3)\right), X_2\right]\right\} \tag{5}$$

$$X''_3 = \mathrm{CSP}\left\{\mathrm{CRC}\left[\mathrm{CBS}(X'_3), X_3, \mathrm{CBS}(X'_2, \mathrm{stride}=2)\right]\right\} \tag{6}$$

$$X''_4 = \mathrm{CSP}\left\{\mathrm{CRC}\left[X'_4, \mathrm{CBS}(X''_3, \mathrm{stride}=2)\right]\right\} \tag{7}$$

where $f_{\mathrm{up}}^{2\uparrow}$ represents the upsampling operation. CBS means 3×3 convolution including batch normalization and SiLU.

Compared with BiFPN, FFM improves the fusion strategy of multiscale feature maps involving reweighting channels. The fusion strategy of BiFPN [29] is between feature maps, which causes different channels have the same weight. In order to strengthen the representation of small object from multiscale features and fully utilize the features of different channels, the proposed CRC reweights the channels of feature map, as shown in the lower half of Fig. 3.

We design three strategies for reweighting channels. The first strategy uses channel attention mechanism similar to SENet [39] or ECANet [53] to reweight channels as formula (8). This strategy is feasible but increases the computational cost and parameter count significantly. The second strategy first concatenates the feature maps and then multiplies the normalized trainable weights with the same number of parameters as the total number of channels, as shown in formula (9). The third strategy further considers the semantic gap between different feature maps, which first reweights the channels within each feature map and then reweights different feature maps, as shown in formula (10)

$$\mathrm{Output} = \mathrm{Attention}(X) \cdot X \tag{8}$$

$$\mathrm{Output} = \sum_j \frac{\omega_j}{\varepsilon + \sum_m \omega_m} \cdot x_j \tag{9}$$

$$\mathrm{Output} = \sum_i \sum_j \frac{\omega_i}{\varepsilon + \sum_k \omega_k} \cdot \frac{\omega_j}{\varepsilon + \sum_{m_i} \omega_{m_i}} \cdot x_j \tag{10}$$

where $\mathrm{Attention}(\cdot)$ represents the channel attention mechanism, such as SENet or ECANet. ω_i represents the trainable weight in the ith feature map. ω_j represents the trainable weight in the jth channel. m_i is the number of channels in the ith feature map. m represents the total number of channels after concatenation. ε is set to 0.0001 to avoid numerical instability. According to the results of ablation experiments in Section IV-D, all the three strategies improve the performance, but the difference between the second and third strategies is not significant. As a result, we select the second strategy in FFM for feature reweighting. The structure of FFM and its channel reweighting strategy optimize the fusion process of multiscale semantic information for small objects, which provides more effective feature maps for subsequent global context modeling.

D. Spatial Context Aware Module (SCAM)

After FEM and FFM, the feature maps have already taken into account local contextual information and have well representation of small object features. Modeling the global relationship between small objects and backgrounds at this stage is more effective than in backbone. Global context information could be used to represent the relationship between pixels cross space, which suppresses useless background and enhances the discrimination between objects and backgrounds. Inspired by GCNet [14] and SCP [38], SCAM consists of three branches. The first branch uses GAP and GMP to integrate global information. The second branch uses a 1×1 convolution to generate linear transform results of the feature map which is named value [54] in Fig. 4. The third branch uses a 1×1 convolution to simplify the multiple of query and key. This convolution is named QK in Fig. 4. Subsequently, the first and third branches are matrix multiplied with the second branch, separately. The obtained two branches represent contextual information cross channels and space, respectively. Finally, the output of SCAM is obtained by using broadcast Hadamard product on these two branches. The structure of SCAM is shown in Fig. 4. In each layer, the pixelwise spatial context can be expressed as follows:

$$Q_i^j = P_i^j + a_i^j \sum_{j=1}^{N_i} \left[\frac{\exp(\omega_{qk} P_i^j)}{\sum_{n=1}^{N_i} \exp(\omega_{qk} P_i^n)} \cdot \omega_v P_i^j\right] \tag{11}$$

$$a_i^j = \frac{\exp\left(\left[\mathrm{avg}(P_i); \max(P_i)\right] P_i^j\right)}{\sum_{n=1}^{N_i} \exp\left(\left[\mathrm{avg}(P_i); \max(P_i)\right] P_i^n\right)} \cdot \omega_v \tag{12}$$

where P_i^j and Q_i^j represent the input and output of the jth pixel in the i-level feature map, respectively. N_i denotes the total number of pixels. ω_{qk} and ω_v are the linear transform matrices for projecting the feature maps, which simplify by 1×1 convolution. $\mathrm{avg}(\cdot)$ and $\max(\cdot)$ perform GAP and GMP, respectively. GAP and GMP can guide feature map to select channels with significant information, which enables SCAM to learn the context information about channel dimensions.

E. Lite-FFCA-YOLO (L-FFCA-YOLO)

A qualified lightweight model needs to strike a balance among parameter count, speed, and accuracy. FasterNet has found that the main reason for low FLOPs of DWConv is its frequent memory redundancy access, which actually leads to the decrease in speed. To alleviate this phenomenon, FasterNet uses PConv, which considers the redundancy in feature maps [51], and applies standard convolution

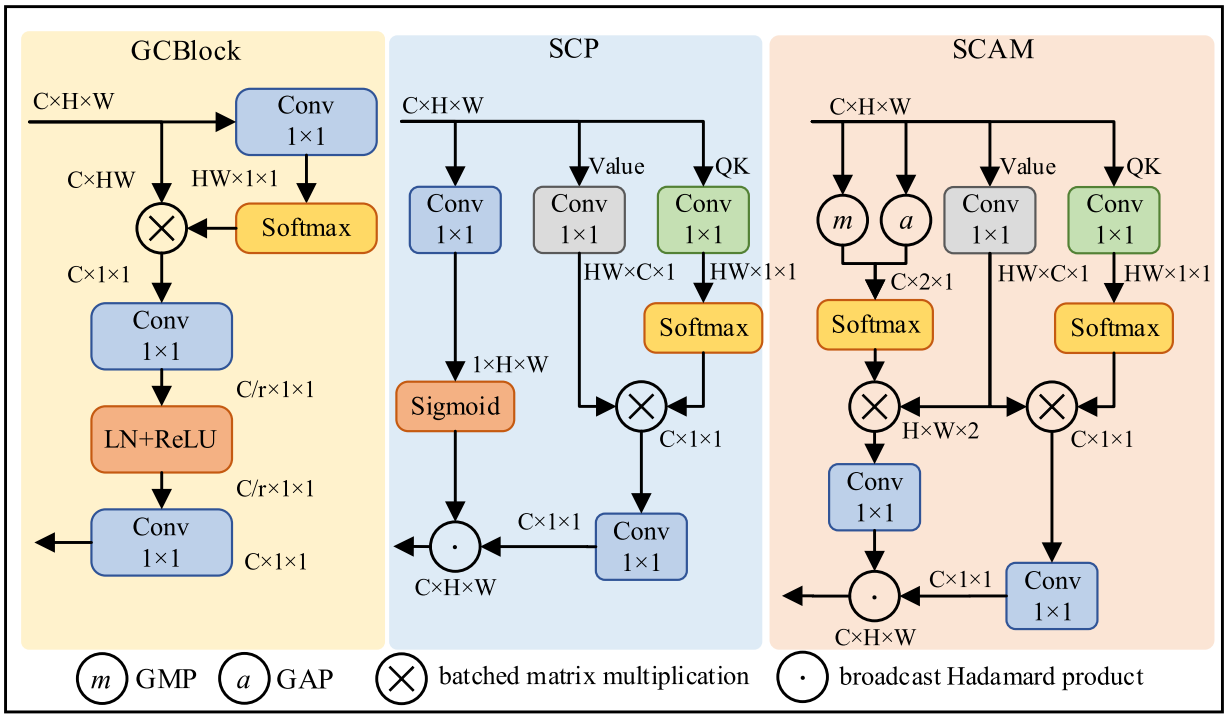


Fig. 4. Structures of GCBlock, SCP, and SCAM.

TABLE I
PARAMETER COUNTS OF FFCA-YOLO AND L-FFCA-YOLO IN BACKBONE

NO.	Module (FFCA-YOLO)	Input	Output	Params	Module (L-FFCA-YOLO)	Input	Output	Params
0	Conv	3	48	5280	Conv	3	48	5280
1	Conv	48	96	41664	Conv	48	96	41664
2	CSPBlock	96	96	111744	CSPFasterBlock	96	96	**61632**
3	FEM	96	96	40968	FEM	96	96	40968
4	Conv	96	192	166272	Conv	96	192	166272
5	CSPBlock	192	192	444672	CSPFasterBlock	192	192	**244224**
6	FEM	192	192	162288	FEM	192	192	162288
7	Conv	192	384	664320	Conv	192	384	664320
8	CSPBlock	384	384	2512896	CSPFasterBlock	384	384	**1310208**
9	SPPF	384	384	369792	SPPF	384	384	369792
Sum				4519896				**3066648**

on only a portion of input channels. The CSPBlock in FFCA-YOLO is reconstructed by combining the FasterBlock in FasterNet, which is named CSPFasterBlock, as shown in Fig. 5.

According to the research results of [51], the number of channels M that using 1×1 convolution is set to 3/4 of the total channels in CSPFasterBlock. Two standard convolutions with channel scaling ratio are set after the PConv. Section IV-E displays the experimental results with different scaling ratios. FasterNet concludes that directly replacing standard convolution with PConv will lead to a serious decline in accuracy. Therefore, we only replace the bottleneck in CSPBlock with FasterBlock, which ensures that the feature information of different layers flows through all channels with little accuracy loss. The parameter counts in the backbone of FFCA-YOLO and L-FFCA-YOLO are presented in Table I, which shows that the backbone of L-FFCA-YOLO has parameters 30% fewer than FFCA-YOLO.

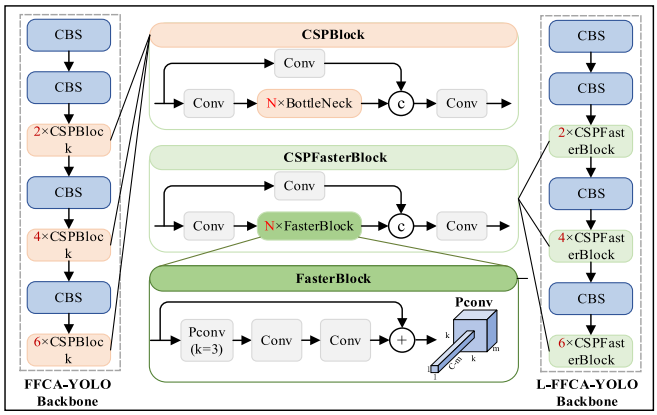


Fig. 5. Backbone structure of L-FFCA-YOLO.

IV. EXPERIMENTAL RESULTS

In this article, small object is defined as an object with size less than 32×32 pixels. The benchmark tests are

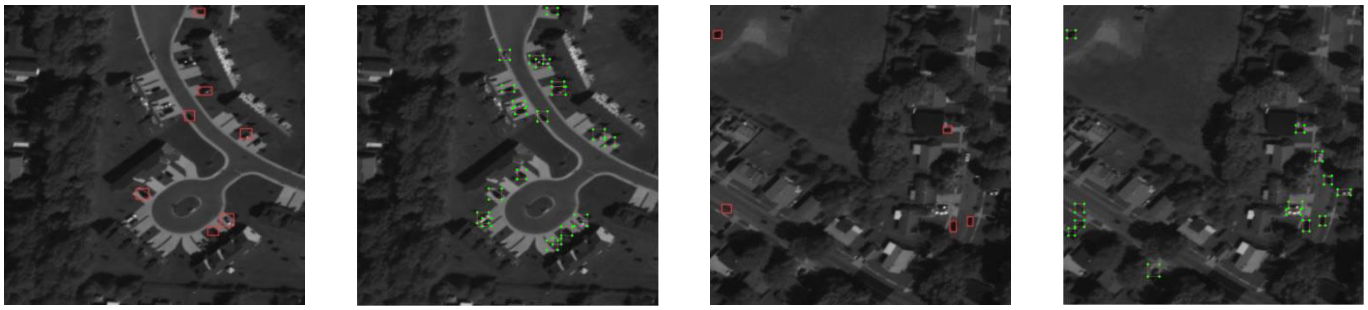


Fig. 6. Ground truth annotation in UNICORN2008 and USOD. The red bounding boxes are the original annotated instances in UNICORN2008, while bounding boxes with green corner points are the manual annotation supplementation for USOD. (a) Original annotation, (b) manual annotation, (c) original annotation, (d) manual annotation.

conducted on two public datasets of small object VEDAI [54] and AI-TOD [55], [56] as well as a self-built dataset (USOD) dedicated to small object detection. YOLOv5 is selected as the benchmark framework, which can be divided into five models with increasing network width and depth: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5m gets the excellent balance between speed and accuracy in the YOLOv5 series of algorithm. Therefore, we use YOLOv5m as the base model and perform improvement and optimization.

A. Experimental Dataset Description

1) VEDAI: Vehicle detection in aerial imagery (VEDAI) dataset [55] consists of cropped images obtained from a larger Utah Automated Geographic Reference Center (AGRC) dataset. In AGRC, each image has about 16 000 × 16 000 pixels, collected from the same altitude, with a resolution of about 12.5 × 12.5 cm per pixel. RGB and IR are two modes of each image in the same scene. We only execute experiments on the RGB version and divide the training and testing sets according to the official given method. We do not consider classes with instances fewer than 50, such as plane, motorcycle, and bus. Our task is to detect eight different classes of objects the same as YOLO-fine [62] and SuperYOLO [63].

2) AI-TOD: AI-TOD [56], [57] is a dataset for tiny object detection in aerial images. Compared with the existing object detection datasets in remote sensing, the average size of the objects in AI-TOD is about 12.8 pixels, which is much smaller than other public datasets. AITOD contains 28 036 aerial images with totaling 700 621 object instances, which are divided into eight classes, including airplane, bridge, storage tank, and so on. We use the training set of 11 214 images and the validation set of 2804 with a total of 14 018 images for training and evaluate the model performance in the test set of 14 018 images according to the official offer.

3) USOD: The existing public datasets [58], [59] contain many medium and large objects, so it is difficult to verify the feature extraction performance of detectors for small objects. Therefore, in order to further verify the detection ability of FFCA-YOLO, unicorn small object dataset (USOD) is built based on UNICORN2008 [60]. UNICORN2008 provides imaging data from photoelectric sensors, whose spatial resolution is about 0.4 m. We used the visible light data of UNICORN 2008 to form USOD by filtering, segmenting, and manually adding annotations for small vehicle objects, as shown in Fig. 6. In addition, UNICORN2008 has SAR images that some of those can be registered with visible light images. In the future, we will add SAR images to USOD for constructing a multimodal version dataset.

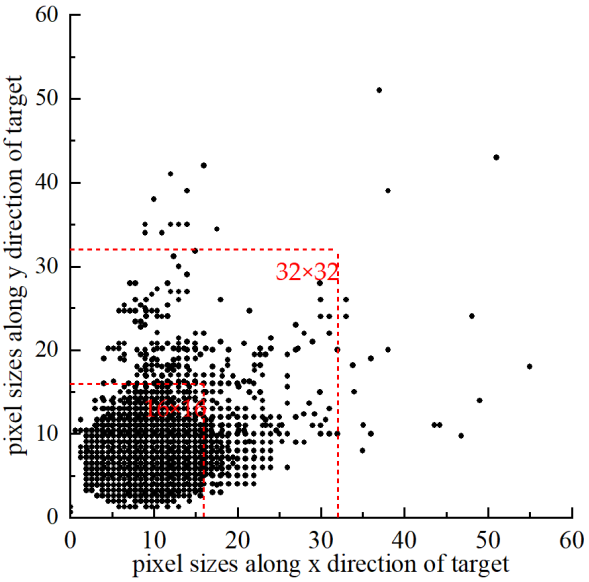


Fig. 7. Distribution of object sizes in USOD.

USOD includes a total of 3000 images containing 43 378 vehicle instances. The ratio of training set to test set is 7:3. As shown in Fig. 7, the proportion of objects with size less than 16 × 16 accounts for 96.3%, and the proportion of objects with size less than 32 × 32 accounts for 99.9%. Fig. 8 shows the distribution of the number of small objects in the training set, which can be seen that small objects are relatively evenly distributed. In summary, USOD can serve as a benchmark dataset for small object detection in remote sensing with the following characteristics.

1) The proportion of small objects in USOD (99.9%) is higher compared with other small object datasets, such as AI-TOD (97.9%).
2) There are many vehicle instances in USOD that are under low illumination and shadow occlusion conditions, which can more effectively validate the performance of models to detect small objects.

TABLE II
COMPARISON EXPERIMENTS FOR FFCA-YOLO IN VEDAI

Methods	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	mAP_{50}	$mAP_{50:95}$	mAP_s
Lightweight CNN[61]	0.747	0.567	0.567	0.361	0.269	0.567	0.227	0.361	0.526	-	-
YOLO-fine[62]	0.767	0.743	0.647	0.634	0.450	0.781	0.700	0.779	0.681	-	-
SuperYOLO(RGB)[63]	0.903	0.826	0.766	0.685	0.538	0.794	0.580	0.703	0.724	-	-
CMAFF(RGB)[64]	0.917	0.859	0.751	0.783	0.333	0.812	0.718	0.622	0.743	-	-
TPH-YOLO[22]	0.840	0.764	0.607	0.629	0.383	0.635	0.237	0.573	0.584	0.338	0.345
YOLOv5m	0.866	0.787	0.724	0.607	0.717	0.797	0.560	0.736	0.723	0.410	0.399
YOLOv8m	0.859	0.81	0.617	0.839	0.56	0.783	0.426	0.595	0.686	0.408	0.401
FFCA-YOLO	0.896	0.857	0.787	0.857	0.486	0.818	0.615	0.67	**0.748**	**0.448**	**0.446**
L-FFCA-YOLO	0.913	0.855	0.728	0.797	0.473	0.79	0.561	0.739	0.733	0.447	0.445

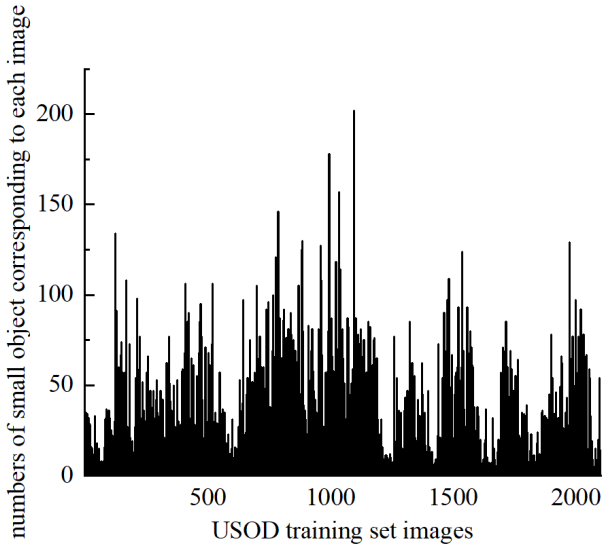


Fig. 8. Distribution of the number of small objects in the training set of USOD.

3) USOD includes a series test sets for validating the robustness of models, considering image degradation factors, including blurring, Gaussian noise, stripe noise, and fog.
4) USOD has the potential to become a multimodal dataset in the future. The data source of USOD is UNICORN2008, which has registered images between visible light data and SAR data.

B. Model Training and Evaluation Metrics

The proposed model was implemented in PyTorch and deployed on a workstation with an NVIDIA 4090 GPU. Stochastic gradient descent (SGD) optimizer was used with initial learning rate 0.01, momentum 0.937, and weight decay 0.0005 to learn the parameters. The batch size during training was set to 32. Normalized Wasserstein distance (NWD) [57] loss is added to the loss function of YOLOv5 as a supplement to the box loss. NWD models the distance between bounding boxes as a Wasserstein distance, which reduced the sensitivity of IOU to small objects. An adjustment weight is introduced for CIOU loss and NWD loss, which is set to 0.5. Mean average precision (mAP) is used as the standard evaluation metric, which can be divided into mAP50, mAP75, mAP50:95, and so on, according to the different IOUs. Here, mAP50 and mAP50:95 are used as the main evaluation metrics. In addition, in order to reflect the detection performance for small objects, mAPs is used as the evaluation metric.

C. Comparisons With Previous Methods

The experimental results of FFCA-YOLO and L-FFCA-YOLO are provided on three datasets: VEDAI, AI-TOD, and USOD. In VEDAI and AI-TOD, we compare our model with current advanced methods and SOTA methods. Fig. 9 shows the detection results of FFCA-YOLO in typical scenarios across various datasets. In USOD dataset, we compare our model with other YOLO models and some classic object detection algorithms.

1) VEDAI: We used 512 × 512 data in VEDAI dataset for training and validating. The results of lightweight CNN [61], YOLO-fine [62], SuperYOLO [63], and CMAFF [64] are compared. Both the original CMAFF and SuperYOLO used multimodal data for training, and we only use their results in training RGB data, which is consistent with our training set. Table II shows that compared with CMAFF, FFCA-YOLO improves by 0.005 in mAP50. Compared with YOLOv5m, FFCA-YOLO improves by 0.025, 0.038, and 0.047 in mAP50, mAP50:95, and mAPs, respectively. Compared to mAP50 and mAP50:95, FFCA-YOLO has a significant improvement in mAPs, which indicates that FFCA-YOLO has a significant advantage over benchmark networks for small object detection in remote sensing.

2) AI-TOD: AI-TOD has a higher proportion of small objects, which better reflects the network's ability in small object detection. The evaluation metrics for AI-TOD dataset are different from other datasets that mAPvt, mAPt, and mAPs are adopted. mAPvt, mAPt, and mAPs represent the mAP for objects with sizes below 8 × 8, 8 × 8 to 16 × 16, and 16 × 16 to 32 × 32, respectively. Table III shows that compared with the SOTA methods, FFCA-YOLO and L-FFCA-YOLO achieve the best performance. In the test set, the mAP50 of FFCA-YOLO reaches 0.617, which is 0.08 higher than the current best model HANet [68]. The mAP50:95, mAPvt, mAPt, and mAPs are increased by 0.056 0.015, 0.027, and 0.045, respectively. The results demonstrate the excellent performance of FFCA-YOLO for small object detection in remote sensing.

3) USOD: Table IV shows the performance of DSSD [69], RefineDet [70], YOLOv3 [19], YOLOv4 [71], YOLOv5m, YOLOv8m, TPH-YOLO [22], and the proposed method in USOD dataset. It can be seen that under the same training

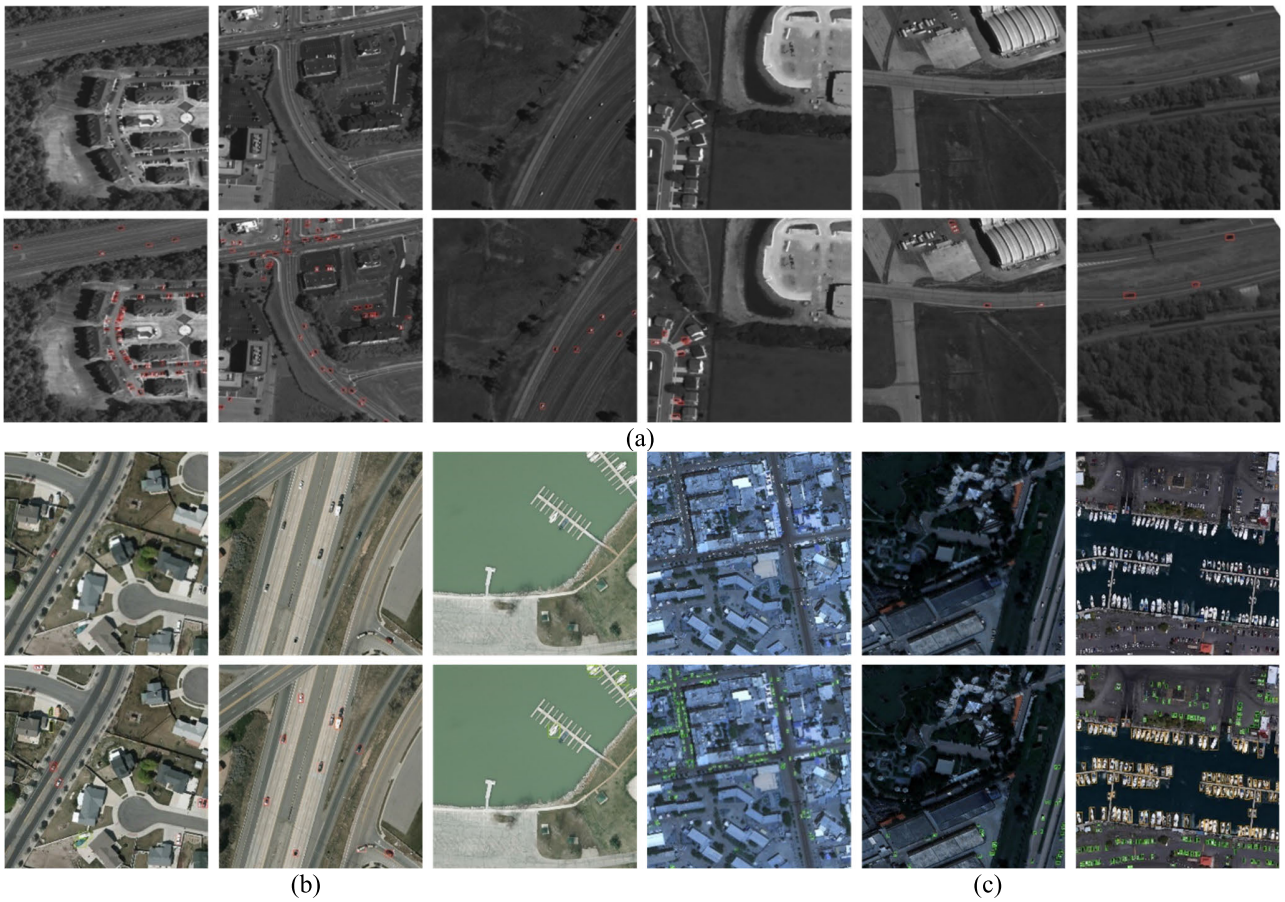


Fig. 9. The detection results of FFCA-YOLO in USOD, VEDAI, and AI-TOD for typical scenarios, such as ports, highways, and buildings. (a) Results in USOD dataset. (b) Results in VEDAI dataset. (c) Results in AI-TOD dataset.

TABLE III
COMPARISON EXPERIMENTS FOR FFCA-YOLO IN AI-TOD

Methods	mAP_{50}	$mAP_{50:95}$	mAP_{vt}	mAP_{t}	mAP_{s}
M-CenterNet[56]	0.407	0.145	0.061	0.150	0.194
Cascade R-CNN[65]	0.308	0.138	0.000	0.106	0.255
DetectoRS[66]	0.328	0.148	0.000	0.108	0.283
DetectoRS + NWD[57]	0.493	0.208	0.064	0.197	0.296
SP-YOLOv8s[67]	0.483	0.227	-	-	-
HANet[68]	0.537	0.221	0.109	0.222	0.273
FFCA-YOLO	**0.617**	**0.277**	**0.126**	**0.249**	**0.318**
L-FFCA-YOLO	**0.583**	**0.255**	**0.117**	**0.232**	**0.301**

TABLE IV
COMPARISON EXPERIMENTS FOR FFCA-YOLO IN USOD

Methods	Backbone	precision	recall	mAP_{50}	$mAP_{50:95}$	mAP_{s}	Para
DSSD[68]	ResNet101	0.645	0.575	0.531	-	-	-
RefineDet[69]	ResNet101	0.881	0.824	0.851	-	-	-
YOLOv3[19]	DarkNet53	0.712	0.694	0.575	-	-	-
YOLOv4[70]	DarkNet53	0.793	0.828	0.778	-	-	-
YOLOv5m	CSPDarkNet53	0.892	0.821	0.873	0.323	0.313	20.85M
YOLOv8m	CSPDarkNet53	0.905	0.822	0.876	0.324	0.314	29.74M
TPH-YOLOv5[22]	CSPDarkNet53	0.910	0.837	0.895	0.321	0.321	45.36M
FFCA-YOLO	CSPDarkNet53	**0.929**	**0.855**	**0.909**	**0.350**	**0.340**	7.12M
L-FFCA-YOLO	CSPDarkNet53	0.928	0.851	0.907	0.349	0.338	5.04M

hyperparameters, FFCA-YOLO has smaller parameter count and higher performance compared with the benchmark methods. L-FFCA-YOLO reduces the parameter count by about 30% compared with FFCA-YOLO (from 7.12 to 5.04M), but showing no significant decline in accuracy metrics. Fig. 10 shows the detection results of YOLOv5m, TPH-YOLO, and

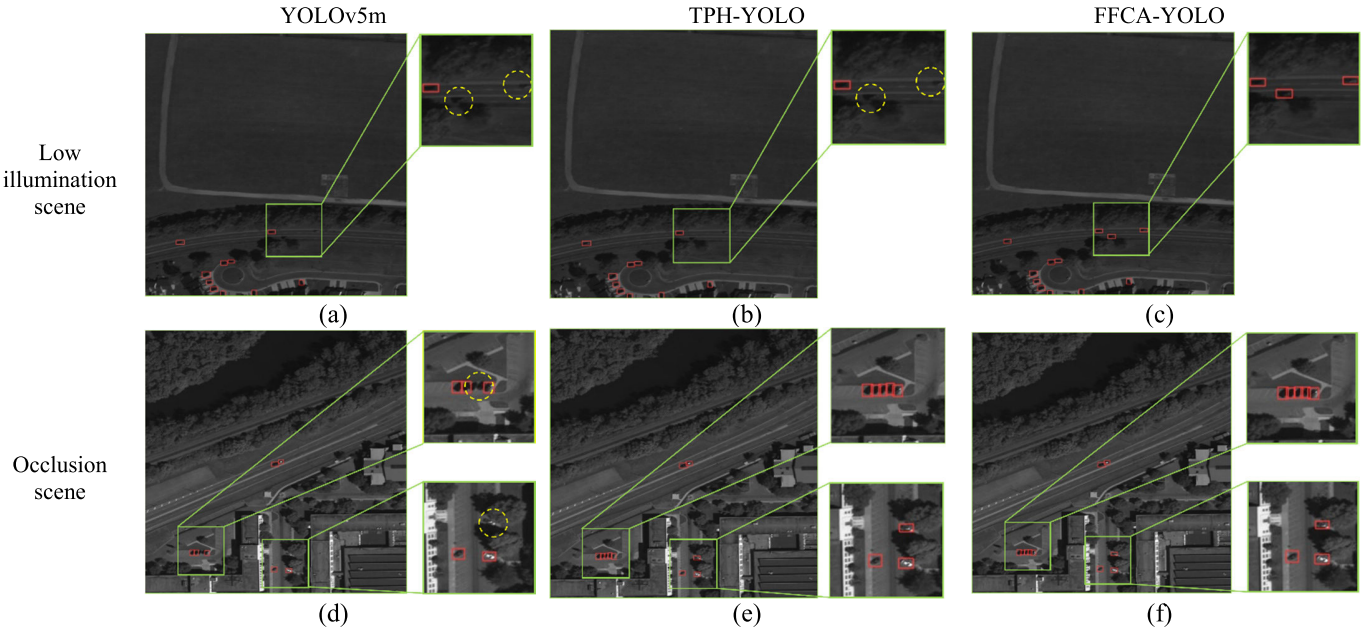


Fig. 10. Detection results of YOLOv5m, TPH-YOLO, and FFCA-YOLO for low illumination and shadow occlusion scenes. The red bounding boxes represent the detection box output by the model, while the yellow circles represent the missed detections.

TABLE V
ABLATION EXPERIMENTS FOR FEM, FFM, AND SCAM IN USOD

FEM	FFM	SCAM	precision	recall	mAP_{50}	$mAP_{50:95}$	mAP_s	Para
×	×	×	0.900	0.826	0.868	0.310	0.303	6.53M
√	×	×	0.926	0.839	0.899	0.343	0.335	6.70M
×	√	×	0.908	0.837	0.876	0.314	0.306	6.54M
×	×	√	0.916	0.828	0.885	0.33	0.321	6.92M
√	√	×	0.928	0.845	0.903	0.345	0.334	6.74M
√	×	√	0.925	0.842	0.901	0.342	0.334	7.09M
×	√	√	0.923	0.851	0.898	0.335	0.324	6.93M
√	√	√	**0.929**	**0.855**	**0.909**	**0.350**	**0.340**	7.12M

FFCA-YOLO in low illumination and shadow occlusion scenes. In low illumination scene, the grayscale values of objects and the background are close to each other causing YOLOv5m and TPH-YOLO to have missed detections. In occlusion scene, one small object is located in the shade of a tree causing YOLOv5m to have missed detection.

D. Ablation Experimental Result

To analyze the importance of each component in FFCA-YOLO, we progressively applied the FEM, FFM, and SCAM in the baseline to verify their effectiveness. The ablation experiment was conducted in USOD dataset. Table V shows the impact of adding or reducing each module on evaluation metrics, where √ represents using the module and × represents not using the module.

1) FEM: As shown in Table V, adding FEM can obviously improve all evaluation metrics, especially in terms of precision (from 0.9 to 0.926) and mAPs (from 0.303 to 0.335). This confirms that FEM makes it easier for the model to distinguish small objects from backgrounds. To further validate this conclusion, we visualize the feature maps before and after FEM in Fig. 11. The brighter color represents that the model pays more attention to that area. Due to FEM enriching the local contextual features, the network has shown good suppression effects on complex backgrounds.

2) FFM: Table V shows that adding FFM can improve all evaluation metrics, especially in terms of recall (from 0.826 to 0.837). In addition, we research on the effects of different neck structures and different fusion strategies of multiscale feature map mentioned in Section III-C, as shown in Table VI. CRC_1, CRC_2, and CRC_3 represent different channel reweighting strategies in formulas (8)–(10), respectively. It can be seen that the performance of CRC_2 and CRC_3 is significantly better in all aspects compared with BiFPN, and the performance difference between CRC_2 and CRC_3 is relatively small (mAP50:95 of CRC_2 is 0.003 higher than that of CRC_3). As a result, CRC_2 is selected as the channel reweighting strategy in FFM.

3) SCAM: Table V shows the performance improvement by adding SCAM. SCAM can improve all evaluation metrics. Table VII shows the comparison among SCAM and some typical baseline methods. SCAM achieves better performance in all evaluation metrics. Fig. 11 shows the impact of SCAM on feature maps. Compared with the feature maps outputted by FEM, the same level feature maps of SCAM further enhance the feature representation of small objects and suppress the

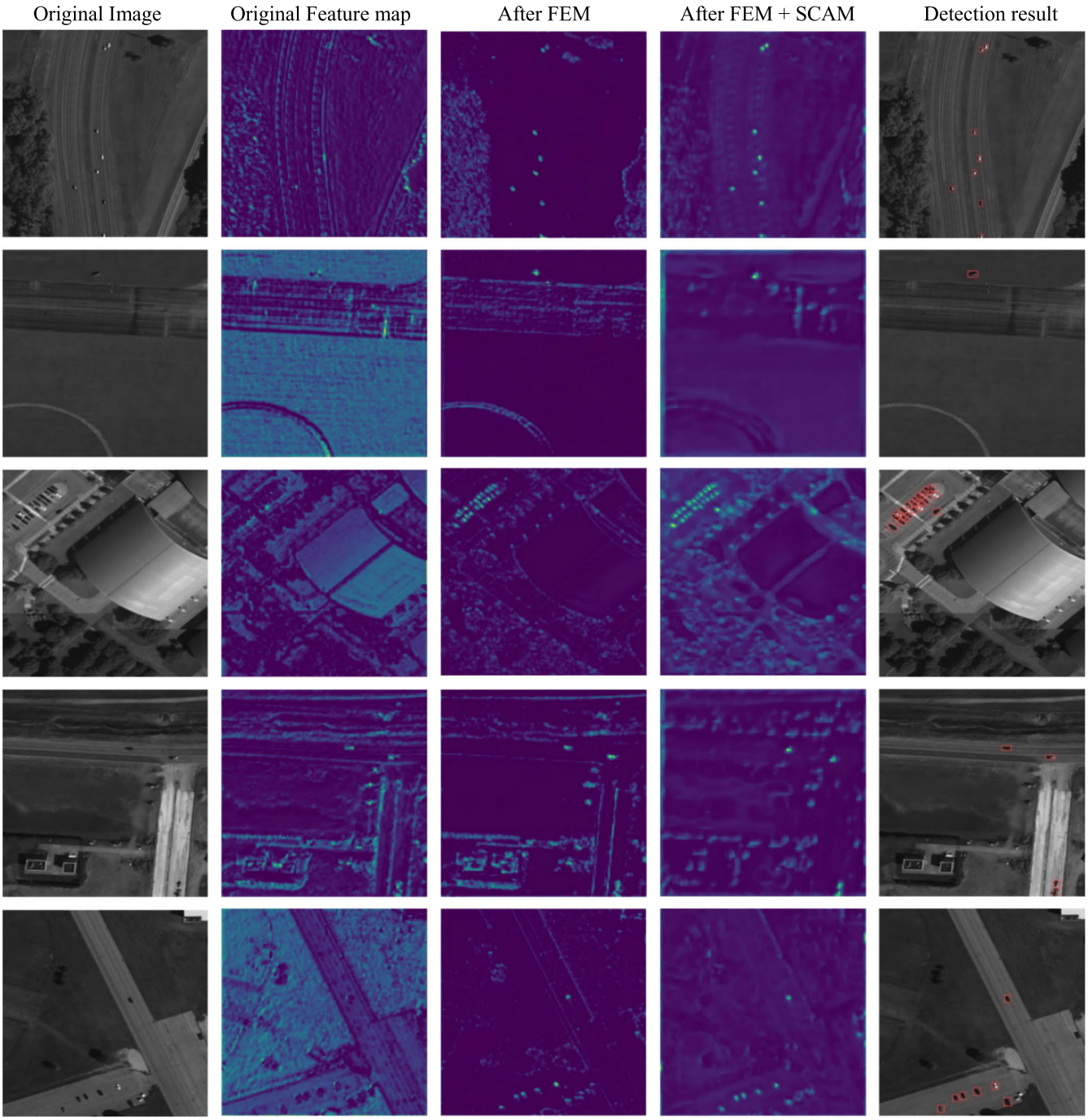


Fig. 11. Influence of FEM and SCAM on feature extraction. The brighter color represents that the model pays more attention to that area.

TABLE VI
COMPARISON EXPERIMENTS FOR FFM IN USOD

Method	precision	recall	mAP_{50}	$mAP_{50:95}$	mAP_S	Para
PANet[26]	0.925	0.842	0.901	0.342	0.334	7.09M
ASFF[28]	0.918	0.840	0.898	0.344	0.333	7.02M
AFPN[72]	0.928	0.853	0.907	0.347	0.338	9.65M
BiFPN(without CRC)	0.927	0.848	0.900	0.341	0.334	7.12M
BiFPN(CRC_1 = SENet[39])	0.926	0.849	0.903	0.342	0.334	7.19M
BiFPN(CRC_1 = ECANet[53])	0.921	0.850	0.897	0.341	0.333	7.13M
BiFPN(CRC_2)	**0.929**	**0.855**	**0.909**	**0.350**	**0.340**	7.12M
BiFPN(CRC_3)	0.927	0.854	0.908	0.347	0.338	7.12M

backgrounds. Through the above analysis of ablation experiments, it can be concluded that the proposed modules FEM, FFM, and SCAM all steadily improve the performance of FFCA-YOLO without any conflicts.

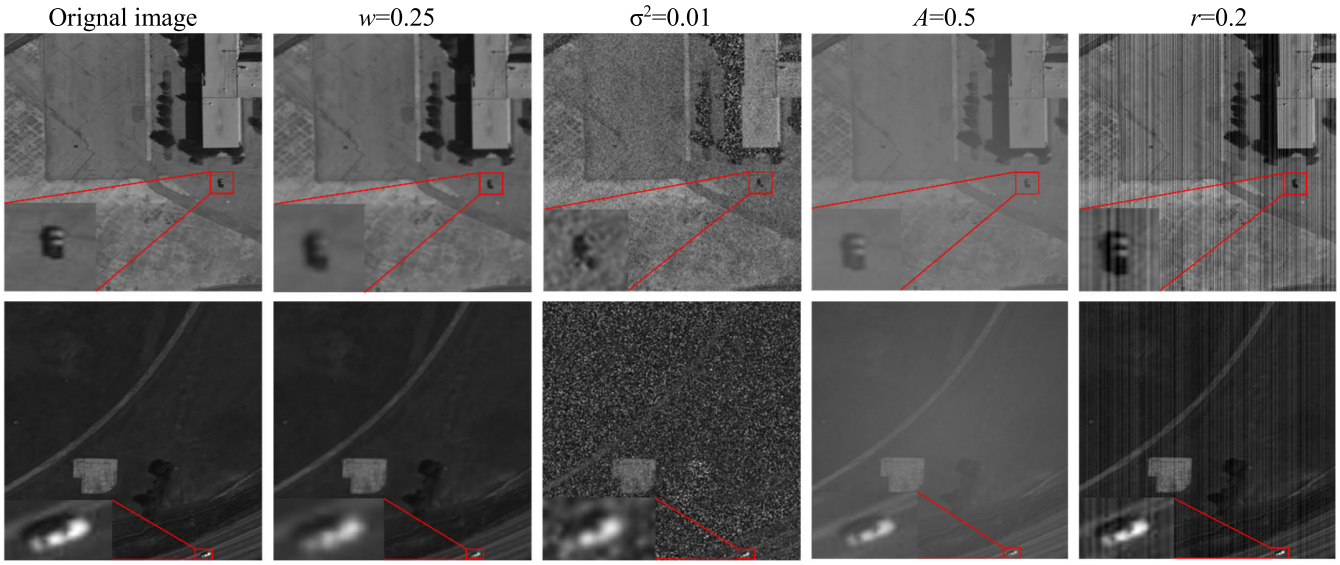


Fig. 12. Simulated degradation images in USOD.

TABLE VII
COMPARISON EXPERIMENTS FOR SCAM IN USOD

Method	precision	recall	mAP_{50}	$mAP_{50:95}$	mAP_S	Para
NLBlock[13]	0.925	0.855	0.905	0.345	0.338	7.52M
SCP[38]	0.925	0.848	0.902	0.344	0.334	7.12M
GCBlock [14]	0.926	0.852	0.907	0.349	0.340	7.12M
SCAM	**0.929**	**0.855**	**0.909**	**0.350**	**0.340**	7.12M

TABLE VIII
ROBUSTNESS EXPERIMENTS FOR FFCA-YOLO AND YOLOV5M IN USOD

Image blurring w	Gaussian noise σ^2	Stripe noise r	Fog A	PSNR	mAP_{50} (FFCA-YOLO)	mAP_{50} (YOLOv5m)	Retrained mAP_{50} (FFCA-YOLO)	Retrained mAP_{50} (YOLOv5m)
-	-	-	-	-	0.909	0.873	0.908	0.856
0.81	-	-	-	44.17	0.907(0.2%↓)	0.872(0.1%↓)	0.904(0.4%↓)	0.837(2.2%↓)
0.64	-	-	-	43.08	0.906(0.3%↓)	0.867(0.6%↓)	0.905(0.3%↓)	0.830(3.0%↓)
0.49	-	-	-	42.32	0.898(1.2%↓)	0.862(1.2%↓)	0.894(1.5%↓)	0.823(3.8%↓)
0.39	-	-	-	41.24	0.893(1.7%↓)	0.855(2.1%↓)	0.883(2.8%↓)	0.814(4.9%↓)
0.25	-	-	-	39.56	0.839(7.7%↓)	0.796(8.8%↓)	0.823(9.3%↓)	0.757(11.5%↓)
-	0.001	-	-	33.54	0.462(49.1%↓)	0.440(49.5%↓)	0.892(1.7%↓)	0.823(3.8%↓)
-	0.005	-	-	24.31	0.021(97.4%↓)	0.041(95.3%↓)	0.805(11.3%↓)	0.739(13.6%↓)
-	0.01	-	-	20.21	0.006(99.3%↓)	0.005(99.4%↓)	0.635(30.0%↓)	0.580(35.7%↓)
-	-	0.05	-	38.14	0.620(31.8%↓)	0.566(35.2%↓)	0.862(5.1%↓)	0.841(1.8%↓)
-	-	0.1	-	32.59	0.212(76.7%↓)	0.162(81.4%↓)	0.824(9.3%↓)	0.805(6.0%↓)
-	-	0.2	-	27.05	0.021(97.7%↓)	0.010(98.9%↓)	0.743(18.2%↓)	0.746(12.9%↓)
-	-	-	0.2	24.91	0.718(21.0%↓)	0.658(24.6%↓)	0.882(2.9%↓)	0.855(0.1%↓)
-	-	-	0.3	22.57	0.574(36.9%↓)	0.542(37.9%↓)	0.854(5.9%↓)	0.845(1.3%↓)
-	-	-	0.4	18.79	0.467(48.6%↓)	0.465(46.7%↓)	0.842(7.3%↓)	0.832(2.8%↓)
-	-	-	0.5	15.81	0.401(55.9%↓)	0.408(53.3%↓)	0.795(12.4%↓)	0.756(11.7%↓)

E. Robustness Experiment

Remote sensing data tend to suffer from various degradation, noise effects, or variabilities in the process of imaging that may cause the aliasing of interested objects and backgrounds, especially when the objects are small. To verify the robustness of FFCA-YOLO under image degradation, we generated a series of test sets that simulating the image degradation in remote sensing based on the research [73]. Each test set has the same original images but different degradation conditions. The degradation types we consider include image blurring, Gaussian noise, stripe noise, and fog. The blurring factor w, the variance of gaussian noise σ^2, and the amplitude factor of the stripe r refer to the article [73]. To generate images with fog, we refer to the model used in [74] and set different atmospheric light parameters A. Fig. 12 shows the degradation results and indicates that image degradation significantly damages the features of small objects. We use peak signal-to-noise ratio (PSNR) to evaluate the quality of degraded images.

TABLE IX
LIGHTWEIGHT EXPERIMENTS FOR L-FFCA-YOLO IN USOD

Method	mAP_{50}	Para	GFLOPS	FPS
CSPBlock	0.909	7.12M	51.2	181
CSPFasterBlock(ratio=2)	0.907	5.04M	37.1	191
CSPFasterBlock(ratio=1)	0.899	4.27M	31.9	207
CSPFasterBlock(ratio=0.5)	0.897	3.89M	29.3	214
GhostBlock	0.889	3.53M	27.3	204
ShuffleBlock	0.832	4.13M	32.9	161

We select FFCA-YOLO and YOLOv5m for robustness testing. The experimental results show that both FFCA-YOLO and YOLOv5m have a certain degree of robustness to the image blurring and fog. FFCA-YOLO has a slightly better effect than YOLOv5m, as shown in Table VIII. Unfortunately, both FFCA-YOLO and YOLOv5m have poor resistance to the impact of gaussian noise and stripe noise, which seriously damage the features of small objects. To alleviate these problems, we add the noise simulation into the data augmentation process and then retrain the models. After retraining, FFCA-YOLO has much better resistance but still unable to deal with images with strong noise. Therefore, we suggest that using image denoising, nonuniformity correction, or other methods to suppress noise before detecting small objects.

F. Lightweight Comparison Experiment

To verify the lightweight effect of L-FFCA-YOLO, CSPFasterBlock is compared with GhostBlock and ShuffleBlock, as shown in Table IX. It can be seen that CSPFasterBlock has a significant performance in mAP50 but with a relatively large number of GFLOPs. That is because GhostNet and ShuffleBlock have more computational redundancy and memory access. Under similar GFLOPs, CSPFasterBlock has faster speed that can optimize speed, accuracy, and memory requirements more effectively. Furthermore, in order to obtain an optimized structure of CSPFasterBlock, different channel scaling ratios are also analyzed in Table IX. It can be found that when the ratio decreases, the mAP50 and parameter count will simultaneously decrease. When the ratio is equal to 2, it has a relatively close performance to FFCA-YOLO.

V. CONCLUSION

In this article, an efficient detector called FFCA-YOLO is designed to detect small objects in remote sensing. Specifically, three lightweight plug-and-play modules (FEM, FFM, and SCAM) are proposed. FEM has multibranch structure to obtain different receptive fields, which fuses local context information of small objects. FFM designs a new feature fusion strategy to reduce the interference of background. SCAM utilizes global pooling to guide global context learning to learn the correlation between channels and reconstructs the correlation between pixels to obtain global context information cross channels and space. In addition, a lite version of FFCA-YOLO named L-FFCA-YOLO uses PConv to reconstruct the backbone and neck. L-FFCA-YOLO has faster speed, smaller parameter scale, and lower computing power requirement but little accuracy loss compared with FFCA-YOLO. The experimental results show that in the two common small object detection datasets VEDAI (RGB) and AITOD, FFCA-YOLO demonstrates the superiority in tasks of small object detection, whose accuracy reaches 0.748 and 0.617 (in terms of mAP50) and exceeds the given SOTA models. In addition, a new small object dataset USOD is constructed, which has a larger proportion of small objects, more scenes with low illumination, and object occlusion, a series of test set under various degradation conditions. The accuracy of FFCA-YOLO on USOD reaches 0.909 (in terms of mAP50), which significantly surpasses other benchmark models, such as YOLOv5m (0.873). Although FFCA-YOLO can achieve good results in small object detection tasks and may have the potential to be applied to real-time processing on board in the future, it still has some limitations.

1) The speed and memory utilization are required to be further optimized before hardware deployment.

2) Currently, the proposed method is only validated on air-based datasets. For space-based remote sensing, the images often have lower resolution, poorer quality, and more complex degradation appearance. Therefore, the effectiveness of our method remains to be further studied and validated.

In the process of research, we find that the ability of the existed deep learning network encounters a bottleneck in small object detection by using only one single-modal data source. In our opinion, multisource combination could enable the detector obtaining more effective feature representations of small objects. As a result, cooperative detection by multiplatform or multiband detection by single platform may be the future development directions in applications for small object detection.

ACKNOWLEDGMENT

The authors would like to express their appreciations to the developers of YOLO and UNICORN2008.

REFERENCES

[1] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image Vis. Comput.*, vol. 97, May 2020, Art. no. 103910.

[2] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.

[3] V. Gagliardi et al., "Satellite remote sensing and non-destructive testing methods for transport infrastructure monitoring: Advances, challenges and perspectives," *Remote Sens.*, vol. 15, no. 2, p. 418, Jan. 2023.

[4] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–22, 2023, Art. no. 5612822, doi: 10.1109/TGRS.2022.3194732.

[5] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 5619513, doi: 10.1109/TGRS.2022.3152250.

[6] F. Zhang, X. Wang, S. Zhou, Y. Wang, and Y. Hou, "Arbitrary-oriented ship detection through center-head point extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5612414, doi: 10.1109/TGRS.2021.3120411.

[7] T. Shi et al., "Feature-enhanced CenterNet for small object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 21, p. 5488, Oct. 2022.

[8] H. Ruan, W. Qian, Z. Zheng, and Y. Peng, "A decoupled semantic-detail learning network for remote sensing object detection in complex backgrounds," *Electronics*, vol. 12, no. 14, p. 3201, Jul. 2023.

[9] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, "Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5786–5795, 2021.

[10] B. Zhang et al., "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022.

[11] B. Vajsova, A. Walczynska, S. Bärisch, P. J. Åstrand, and S. Hain, "New sensors benchmark report on WorldView-4," Publications Office Eur. Union, Luxembourg, U.K., Tech. Rep. EUR 28761 EN, 2017.

[12] R. Trautner and R. Vitulli, "Ongoing developments of future payload data processing platforms at ESA," in *Proc. On-Board Payload Data Compress. Workshop (OBPDC)*, 2010.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[14] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[15] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Global context networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6881–6895, Jun. 2023.

[16] S. Q. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[20] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[21] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[22] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[23] M. Wang et al., "FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection," *J. Vis. Commun. Image Represent.*, vol. 90, Feb. 2023, Art. no. 103752.

[24] L. Shen, B. Lang, and Z. Song, "CA-YOLO: Model optimization for remote sensing image object detection," *IEEE Access*, vol. 11, pp. 64769–64781, 2023.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[26] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[27] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[28] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[29] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[30] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving multi-scale feature learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12595–12604.

[31] Z. Liu, G. Gao, L. Sun, and Z. Fang, "HRDNet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[32] G. Cheng et al., "Feature enhancement network for object detection in optical remote sensing images," *J. Remote Sens.*, vol. 1, p. 14, 2021, doi: 10.34133/2021/9805389.

[33] K. Zhang and H. Shen, "Multi-stage feature enhancement pyramid network for detecting objects in optical remote sensing images," *Remote Sens.*, vol. 14, no. 3, p. 579, Jan. 2022.

[34] R. Liu et al., "RAANet: A residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 13, p. 3109, Jun. 2022.

[35] Y. Li, Z. Cheng, C. Wang, J. Zhao, and L. Huang, "RCCT-ASPPNet: Dual-encoder remote image segmentation based on transformer and ASPP," *Remote Sens.*, vol. 15, no. 2, p. 379, Jan. 2023.

[36] W. Chen, S. Ouyang, W. Tong, X. Li, X. Zheng, and L. Wang, "GCSANet: A global context spatial attention deep learning network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1150–1162, 2022.

[37] Y. Zhou et al., "BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, Art. no. 5618617, doi: 10.1109/TGRS.2022.3152575.

[38] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo, and C. Wen Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," 2021, *arXiv:2111.11057*.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[40] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*., 2018, pp. 3–19.

[41] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2736–2744.

[42] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1389–1397.

[43] S. Guo, Y. Wang, Q. Li, and J. Yan, "DMCP: Differentiable Markov channel pruning for neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1536–1544.

[44] J. Chang, Y. Lu, P. Xue, Y. Xu, and Z. Wei, "Automatic channel pruning via clustering and swarm intelligence optimization for CNN," *Appl. Intell.*, vol. 52, pp. 17751–17771, Apr. 2022.

[45] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[46] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[47] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1580–1589.

[48] L. Huyan et al., "A lightweight object detection framework for remote sensing images," *Remote Sens.*, vol. 13, no. 4, p. 683, Feb. 2021.

[49] J. Yi, Z. Shen, F. Chen, Y. Zhao, S. Xiao, and W. Zhou, "A lightweight multiscale feature fusion network for remote sensing object counting," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, Art. no. 5902113, doi: 10.1109/TGRS.2023.3238185.

[50] J. Liu, R. Liu, K. Ren, X. Li, J. Xiang, and S. Qiu, "High-performance object detection for optical remote sensing images with lightweight convolutional neural networks," in *Proc. IEEE 22nd Int. Conf. High Perform. Comput. Commun., IEEE 18th Int. Conf. Smart City, IEEE 6th Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Dec. 2020, pp. 585–592.

[51] J. Chen et al., "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12021–12031.

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

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[54] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[55] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016.

[56] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny object detection in aerial images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3791–3798.

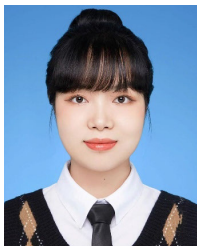
[57] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.
[58] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[59] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[60] L. Colin et al. (2019). *Unified Coincident Optical and Radar for Recognition (UNICORN) 2008 Dataset*. [Online]. Available: https://github.com/AFRL-RY/data-unicorn-2008
[61] M. A. Momin, M. H. Junos, A. S. M. Khairuddin, and M. S. A. Talip, "Lightweight CNN model: Automated vehicle detection in aerial images," *Signal, Image Video Process.*, vol. 17, no. 4, pp. 1209–1217, Jun. 2023.
[62] M.-T. Pham, L. Courtrai, C. Friguet, S. Lefèvre, and A. Baussard, "YOLO-Fine: One-stage detector of small objects under various backgrounds in remote sensing images," *Remote Sens.*, vol. 12, no. 15, p. 2501, Aug. 2020.
[63] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, Art. no. 5605415, doi: 10.1109/TGRS.2023.3258666.
[64] F. Qingyun and W. Zhaokui, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108786.
[65] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[66] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10213–10224.
[67] M. Ma and H. Pang, "SP-YOLOv8s: An improved YOLOv8s model for remote sensing image tiny object detection," *Appl. Sci.*, vol. 13, no. 14, p. 8161, Jul. 2023.
[68] G. Guo, P. Chen, X. Yu, Z. Han, Q. Ye, and S. Gao, "Save the tiny, save the all: Hierarchical activation network for tiny object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 221–234, Jan. 2024, doi: 10.1109/TCSVT.2023.3284161.
[69] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[70] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[71] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[72] G. Yang, J. Lei, Z. Zhu, S. Cheng, Z. Feng, and R. Liang, "AFPN: Asymptotic feature pyramid network for object detection," 2023, *arXiv:2306.15988*.
[73] C. Li, Z. Li, X. Liu, and S. Li, "The influence of image degradation on hyperspectral image classification," *Remote Sens.*, vol. 14, no. 20, p. 5199, Oct. 2022.
[74] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2010.

Yin Zhang received the B.Sc. degree from Jilin University, Changchun, China, in 2009, and the M.Sc. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2011 and 2016, respectively.

He is currently an Associate Professor with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His main research interests include simulating and processing photoelectric detection information.

Mu Ye received the B.Sc. and M.Sc. degrees from the Shanghai University of Engineering Science, in 2019 and 2022, respectively. He is currently pursuing the D.Eng. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His main research interests include space-based object detection and signal processing.



Guiyi Zhu received the B.Sc. degree from Xidian University, Xi'an, China, in 2015. She is currently pursuing the M.S. degree with the Nanjing University of Aeronautics and Astronautics, Nanjing, China.

Her main research interests include object detection and classification.

Yong Liu received the B.Sc. and M.Sc. degrees from Air Force Aviation University, Changchun, China, in 2012 and 2014, respectively, and the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2018.

He is currently a Research Assistant with the National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, China. His main research interests include remote sensing data processing and information fusion.

Pengyu Guo received the master's degree in computer science and technology from the National University of Defense Technology, Changsha, China, in 2008, and the Ph.D. degree in aerospace science and technology from the National University of Defense Technology, in 2015.

He has experience in working for the China Xi'an Satellite Control Center, Xi'an, China. He is currently an Associate Research Fellow with the National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing, China. His current research focus is to devise algorithms based on computer vision and machine learning to enable unmanned platform's imaging systems for detection, tracking, recognition, and relative pose estimation.

Junhua Yan received the B.Sc., M.Sc., and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1993, 2001, and 2004, respectively.

She is currently a Professor with the Nanjing University of Aeronautics and Astronautics. Her main research interests include image quality assessment, multisource information fusion, object detection, tracking, and recognition.