ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery

Yanfeng Liu, *Student Member, IEEE*, Qiang Li, *Graduate Student Member, IEEE*, Yuan Yuan, *Senior Member, IEEE*, Qian Du, *Fellow, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Benefiting from the development of convolutional neural networks (CNNs), many excellent algorithms for object detection have been presented. Remote sensing object detection (RSOD) is a challenging task mainly due to: 1) complicated background of remote sensing images (RSIs) and 2) extremely imbalanced scale and sparsity distribution of remote sensing objects. Existing methods cannot effectively solve these problems with excellent detection accuracy and rapid speed. To address these issues, we propose an adaptive balanced network (ABNet) in this article. First, we design an enhanced effective channel attention (EECA) mechanism to improve the feature representation ability of the backbone, which can alleviate the obstacles of complex background on foreground objects. Then, to combine multiscale features adaptively in different channels and spatial positions, an adaptive feature pyramid network (AFPN) is designed to capture more discriminative features. Furthermore, considering that the original FPN ignores rich deep-level features, a context enhancement module (CEM) is proposed to exploit abundant semantic information for multiscale object detection. Experimental results on three public datasets demonstrate that our approach exhibits superior performance over baseline by only introducing less than 1.5M extra parameters.**

***Index Terms*— Adaptive feature pyramid, context exploitation, local cross-channel attention, multiscale object detection, remote sensing image (RSI).**

I. Introduction

With the development of aerial technology, the acquisitions and applications of remote sensing images (RSIs) have become more diverse [1]–[3]. Remote sensing object detection (RSOD) is one of the hot research topics in the field of RSIs analysis. It not only locates the object regions of interest in RSIs but also categorizes the classes of multiobjects, which has been widely used in hazard response [4], urban monitoring [5], traffic control [6], and so on. Although many algorithms have been proposed for RSOD, especially for large-scale RSIs, this task still remains challenges mainly due to complex scenes and multiscale objects.

Different from natural scene images, RSIs are commonly captured from satellites with wide views, which leads to the large-scale images and background clutter [7]. Furthermore, objects in different RSIs are in various scales by reason of the variation in image acquisition altitudes [8]. Besides, certain categories of objects are usually distributed densely in RSIs, such as ships and vehicles [9]. The above issues are the main obstacles for object detection in RSIs, which makes most algorithms for natural images not adapted to RSIs well.

Most RSOD algorithms based on convolutional neural networks (CNNs) are motivated by corresponding methods for natural images. The mainstream object detection approaches can be roughly divided into two types: two-stage and one-stage. The former defines the task as a step-by-step refining process (regions extraction and bounding boxes classification), while the latter performs a one-step process. Faster RCNN [10] is a representative two-stage method that implements the first end-to-end network for general object detection. Its main innovation is to design a region proposal network (RPN) to gather proposals instead of a sliding window. The typical one-stage methods mainly include YOLO [11]–[13], RetinaNet [14], and so on. For example, YOLO applies a single network to the input image and divides the image into several cells. Then, it outputs the predicted bounding boxes (b-boxes) and categories probabilities of each region directly. However, these algorithms are not good at dealing with multiscale objects. For instance, Faster RCNN [10] and YOLOv1, v2 [11], [12] only make predictions on the last layer of features. Based on this deficiency, feature pyramid networks (FPNs) [15]–[17] are adopted to handle multiscale features for detection. A number of improved FPNs have been widely studied [18]–[20] thereafter. Nevertheless, FPNs only address multiscale imbalance at the feature level, which cannot settle other imbalance problems. Therefore, Pang *et al.* [21] propose balanced sampling and balanced smooth L1 loss to restrain sample and objective level imbalance respectively. Besides, Chen *et al.* [22] propose an overlap sampler to select examples and enable training to solve the imbalance of sampling. A neoteric loss function [23] is designed during the distillation process to attract positive pixels and reduce area imbalance of foreground and background. These models for natural scene object detection promote the corresponding development in remote sensing fields.

Manuscript received August 23, 2021; revised October 22, 2021; accepted December 4, 2021. Date of publication December 8, 2021; date of current version February 25, 2022. This work was supported by the National Natural Science Foundation of China under Grant U21B2041, Grant U1864204, Grant 61632018, and Grant 61825603. *(Corresponding author: Qi Wang.)*

Yanfeng Liu and Qiang Li are with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liuyanfeng99@gmail.com; liqmges@gmail.com).

Yuan Yuan and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: y.yuan1.ieee@gmail.com; crabwq@gmail.com).

Qian Du is with the Department of Electronic and Computer Engineering, Mississippi State University, Starkville, MS 39759 USA (e-mail: du@ece.msstate.edu).

Digital Object Identifier 10.1109/TGRS.2021.3133956

Considering the various characteristics of RSIs, plenty of improved algorithms (based on natural scene object detection methods) have been applied to remote sensing areas, such as [24]–[26]. To alleviate the object confusion caused by complex background in RSIs, RFN [27] embeds squeeze-and-excitation (SE) blocks [28] into detector for feature selection. Wang *et al.* [29] propose FRPNet that adds the convolutional block attention module (CBAM) [30] into FPN to exploit global information from complicated scenes. Beyond that, there are several approaches to design other spatial or channel attention mechanisms [31]–[33]. To detect multiscale objects accurately, Guo *et al.* [26] adopt balanced FPN to handle with imbalance of feature level for aerial ship detection. FMSSD [34] designs a spatial pyramid with several parallel dilated convolutional layers to enlarge receptive fields, which involves a large amount of calculation. CAD-Net [32] forms feature pyramid by the spatial-and-scale-aware attention module to explore more informative region proposals at different scales. To detect densely packed objects in RSIs, SCRDet [35] introduces a supervised multidimensional attention network to inhibit the adverse effects of background noise. ClusDet [36] develops a cluster proposal subnet to predict cluster regions by a supervised process. However, none of the existing public remote sensing datasets provides ground truth for clusters.

Nonetheless, most of these approaches cannot perform very well in sophisticated scenes, especially for multiscale and dense objects in large-scale RSIs. For example, the attention mechanisms [27], [29], [32], [33], [35] primarily designed with fully connected layers cannot be applied to RSIs well since they are not efficient to integrate into CNNs. On the one hand, these methods increase the running time with costly computation. On the other hand, they make it difficult to fine-tune the networks. Meanwhile, although various feature pyramids, such as [34], [37], [38], have improved the detection performance of RSIs to a certain extent, their structures are still intricate with heavy computation. Besides, it is also a challenge to detect clustered objects accurately in RSIs [8], [9]. Thus, how to handle these problems for RSOD still needs more research efforts.

To address the abovementioned issues, we propose an adaptive balanced network (ABNet) with a series of components to improve detection accuracy and maintain superior running speed. Initially, to relieve the complicated background of large-scale RSIs, we design a portable mechanism named enhanced effective channel attention (EECA) to capture local cross-channel correlation. Then, an adaptive feature pyramid network (AFPN) is proposed, which first integrates multiscale feature maps by adaptive pooling. After that, we present a novel selective refined module (SRM) to reconstruct AFPN. Furthermore, a context enhancement module (CEM) is designed to mitigate the lack of contextual information in integrated features and construct a multiscale pyramid network for detection. The multilevel RPN [15] is utilized to create and identify proposal candidates according to the multiscale pyramid features. Finally, balanced L1 loss [21] is adopted to train the detector steadily and accurately. Extensive experiments are conducted to evaluate the effectiveness of the proposed detector on three representative datasets. The contributions of this article are summarized as follows.

1) Aiming at large-scale RSIs with a complex background, we design the EECA mechanism to extract fine-grained features. It is lighter to deploy in deep CNNs with bringing few parameters and achieving better performance than SE [28], CBAM [30], and ECA [39].
2) To solve multiscale and dense object detection efficiently, a novel pyramid network (AFPN) is presented. It contains SRM to refine different feature maps selectively.
3) We propose the CEM to address the low efficiency of rich channel information of the backbone. Considering the aliasing effect caused by CEM, an adaptive spatial fusion module (ASFM) is introduced to combine contextual features adaptively.

Experiments show that our detector increases 3.41% and 3.26% mAP on the NWPU VHR-10 [7] dataset and the RSOD [40] dataset, respectively. Meanwhile, it achieves 72.8% mAP on the DIOR [8] dataset without bells and whistles. The remainder of this article is organized as follows. We briefly review the related studies in Section II and describe the methodology in Section III. The experiments and model analysis are provided in Section IV. Section V draws a conclusion.

II. Related Work

A. Channel Attention Mechanisms

The attention module in deep learning is first proposed in the machine translation area [41]. In recent years, attention mechanisms of CNNs have attracted great attention in computer vision [28], [30], [39], [42]. It is straightforward to understand the basic principle of channel attention mechanisms. Their objective is to attach importance weights for different channels and make CNNs focus on discriminative feature maps, thus improving performance.

Due to the simplicity and effectiveness of channel attention, learning various important weights of different channels has become a popular and powerful tool in the computer vision community. The most representative channel attention mechanisms are summarized as follows.

1) SENet [28] focuses on the relationship between different channels by merging the SE modules into ResNet [43]. Global average pooling (GAP) is used to estimate channel weights, and multilayer perceptron (MLP) is adopted to accomplish nonlinear mapping of weights. SENet learns to acquire the importance of all channels and adds or suppresses features adaptively. Although SENet improves feature extraction capabilities, the MLP structures prompt the network more overweight and not suitable for fine-tuning in remote sensing tasks [27].
2) CBAM[1] [30] employs a global max-pooling (GMP) and a GAP to output channel weights and then applies a shared MLP to learn attention maps. Similar to SENet [28], CBAM also includes a large number of parameters due to numerous fully connected layers.

[1]For a fair comparison of channel modeling capabilities, the CBAM adopted in this article does not include spatial attention modules.

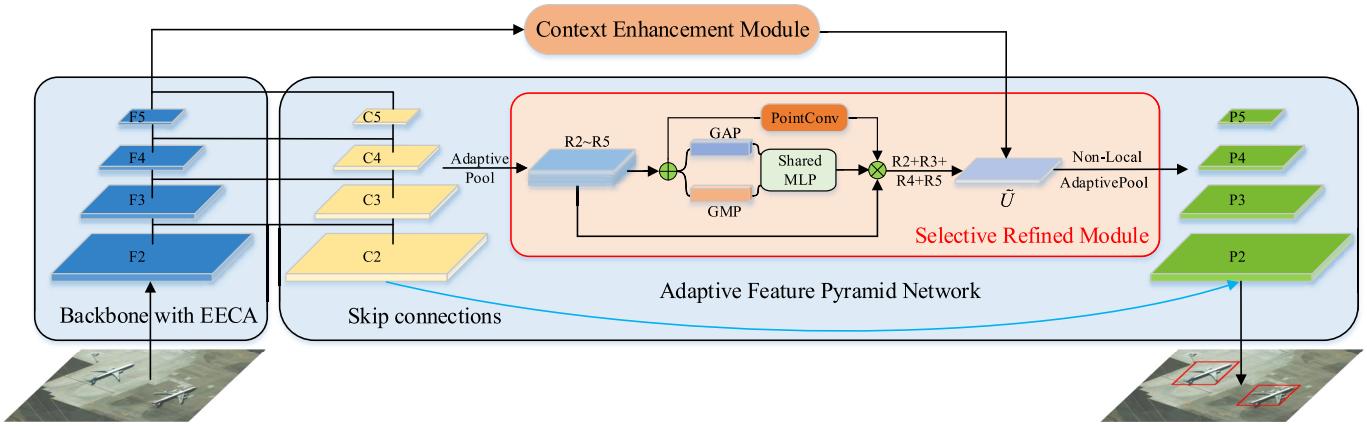

Fig. 1. Illustration of our proposed ABNet. For an input RSI, ABNet first applies ResNet50 backbone (modified by EECA) to extract multiscale feature maps $\{F_2, F_3, F_4, F_5\}$. Then, AFPN utilizes its selective fusion strategy to produce the aggregated feature map $\widetilde{U}$. After that, ABNet uses the context enhanced module to ameliorate $\widetilde{U}$ and obtain multiscale features $\{P_2, P_3, P_4, P_5\}$ for detection.

3) ECANet [39] deploys a 1-D convolutional layer with the kernel of k (5 or 7) to form a local channel-dependent module to extract attention weights. The experiments show that it is better than SE and CBAM, and it can reduce the redundancy of fully connected layers simultaneously.

Notoriously, the above mechanisms are all designed for natural images. However, they result in suboptimal performance in RSIs due to the inconsistent data distribution. Besides, the huge amount of extra parameters (as illustrated in Table I, e.g., SE and CBAM) increases the difficulty of fine-tuning the networks in aerial tasks. In Section III, we design a light mechanism named EECA according to the characteristics of RSIs. Importantly, EECA is not only more lightweight than SE [28] and CBAM [30] but also superior to ECA [39] for RSIs.

B. Feature Pyramids for Object Detection

FPN is first presented in [15]. It is an elegant method to solve multiscale object detection by using upsampling and elementwise summation to coalesce feature maps with different scales. This strategy is also adopted by YOLOv3 [13] and RetinaNet [14]. After that, many advanced FPNs have been proposed, such as [16]–[18]. DFPN [16] is composed of a global attention model (SE blocks [28]) and a local reconfiguration model (residual blocks), which can interact with features across locations and scales. PAFPN [17] adds a bottom-up path to better integrate multiscale feature maps. AugFPN [18] puts forward three modules to solve three shortcomings of FPN and achieves obvious gain in detection performance. In summary, these FPNs introduce heavy modules to improve detection accuracy but cannot maintain running speed.

In the field of RSOD, FPN is still one of the most popular detection paradigms. Recently, the published algorithms, such as CAD-Net [32], GL-Net [33], FMSSD [34], CANet [37], SB-MSN [38], FSoD-Net [44], CF2PN [45], and ASSD [46], also propose a variety of FPNs to detect diverse remote sensing objects. Unfortunately, they cannot solve both the problem of detecting clustered objects and rapid detection due to feature confusion or intricate structures. Different from them, we propose an improved FPN with adding very few parameters to keep considerable detection speed, which can detect multiscale and dense objects efficiently.

C. Context Exploitation

The context information in object detection describes the specific relationship between objects and scenarios. Several papers show the significance of using context for object detection [18], [20], [47]. Therefore, it is more crucial to extract context information for RSIs with complex backgrounds. Li *et al.* [48] propose a local-contextual feature fusion module to build powerful joint representations for RSOD. Experiments testify that this module can integrate local features with global information efficiently. Recently, various context extraction modules have been presented in several optical RSOD algorithms. For example, CAD-Net [32] deploys a global context network and a pyramid local context network to learn the global and local scenes of the objects respectively. GLNet [33] and FSoD-Net [44] also create context modules to deal with sophisticated scenes issue and improve detection performance.

However, none of the above methods takes into account the reduction of channels in deep-level features of the backbone. In this article, we propose a context module named CEM to address the low efficiency of rich channel information in the backbone and exploit context features concurrently. More importantly, our module is lightweight and effective for RSOD, which is validated by experiments in Section IV.

III. Methodology

The framework of ABNet is summarized in Fig. 1. It applies EECA to modify the backbone network [43] and improve feature extraction capability for the input images. Then, it accommodates the original feature pyramid with SRM to achieve AFPN. Finally, CEM is adopted to enrich features of

TABLE I
PARAMS COMPARISON OF FOUR ATTENTION MECHANISMS IN RESNET50

stage name	channels	blocks	SE or CBAM	ECA	EECA
conv2x	256	3	2×256×16	1×5	2×9
conv3x	512	4	2×512×32	1×5	2×9
conv4x	1024	6	2×1024×64	1×7	2×11
conv5x	2048	3	2×2048×128	1×7	2×11
total parameters			2514944	98	324

Assuming the input vector is $X \in \mathbb{R}^{C\times 1\times 1}$, for a MLP, the total number of params is $2 \times C \times C/r$, where $r = 16$ (refer to [28]); for a 1-D convolution, the total number of params is k, where k is the kernel size of 1-D convolution.
For example, the total params of ECA is 5×3 + 5×4 + 7×6 + 7×3 = 98.

AFPN for multiscale object detection. The details of the three components and loss function are presented in the following.

A. Enhanced Effective Channel Attention (EECA)

Channel attention mechanisms are efficient to help the backbone extract more informative features. The state-of-the-art channel attention mechanisms are all designed for natural images [28], [30], [39]. However, their direct application to RSIs fails because of the differences in data distribution (as illustrated in Table II). To restrain the disturbance of complicated background, we design an EECA mechanism inspired by ECA [39]. Its structure is shown in Fig. 2. Different from ECA, our EECA utilizes two 1-D convolutional layers to capture nonlinear local cross-channel interaction. In addition, EECA uses GMP and GAP to obtain channel weights since it can take the most significant knowledge of each channel into account.

Taking an intermediate tensor $X \in \mathbb{R}^{C\times H\times W}$ as input, EECA deploys both GAP and GMP to generate two global spatial context maps: $X_{\text{avg}}, X_{\max} \in \mathbb{R}^{C\times 1\times 1}$, which denote average-pooled features and max-pooled features, respectively. GAP and GMP can be expressed as

$$X_{\text{avg}} = \text{GAP}(X) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j} \tag{1}$$

$$X_{\max} = \text{GMP}(X) = \max \sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j}. \tag{2}$$

Then, both X_{avg} and $X_{\max}$ are input to a shared block to generate channel attention map M. Specifically, the shared block is composed of two 1-D convolutional layers and an ReLU layer for local cross-channel correlation. EECA merges the output feature maps by using elementwise summation after exploiting each descriptor. Thus, the channel attention $M(X)$ is computed as

$$M(X) = \sigma(\text{C}_2(\text{RL}(\text{C}_1(X_{\text{avg}}))) + \text{C}_2(\text{RL}(\text{C}_1(X_{\max})))) \tag{3}$$

where $\sigma(\cdot)$ indicates the Sigmoid function, $\text{RL}(\cdot)$ denotes the ReLU function, and C_1 and C_2 indicate the first and second 1-D convolutional layers, respectively. EECA can obtain the final refined feature map $\widetilde{X}$ via elementwise multiplication of $M(X)$ and X, i.e.,

$$\widetilde{X} = X \otimes M(X) \tag{4}$$

where $\otimes$ represents elementwise multiplication.

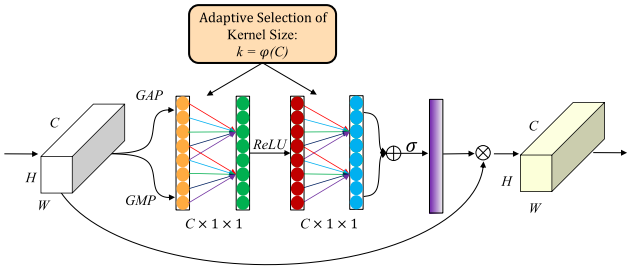

Fig. 2. Illustration of EECA. Given the aggregated features obtained by GAP and GMP, EECA generates channel attention by performing two fast 1-D convolutions of size k with ReLU and Sigmoid functions.

As for the setting of kernel size of 1-D convolutional layers, we refer to the nonlinear mapping function in ECA [39]. It is assumed that the size of kernel k is positively correlated with the number of channels C. Here, it can be formulated as

$$k = \varphi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{\text{odd}} = |\log_2(C)|_{\text{odd}} \tag{5}$$

where $|t|_{\text{odd}}$ denotes the nearest odd number of t. If t is an even number, then $|t|_{\text{odd}} = t + 1$; otherwise, $|t|_{\text{odd}} = t$. Unlike ECA [39], we set γ to 1 because it is of benefit to RSIs by focusing on a larger cross-channel interaction. For simplicity, b is set to 0.

Through integrating EECA modules into residual blocks of ResNet50 [43], an enhanced feature extraction network is reached. The differences of parameters between four attention mechanisms are clearly compared in Table I. It is worth mentioning that our EECA only adds 324 parameters (less than 1% of SE [28]) to merge with ResNet50, which is perfect for model fine-tuning.

EECA pays attention to more spatial context information efficiently than ECA [39]. Compared with SE [28] and CBAM [30], EECA applies local convolutional operations instead of fully connected layers, which greatly reduces the number of parameters and computational complexity. Experiments in Section IV reveal the superiority of EECA for RSIs.

B. Adaptive Feature Pyramid Network (AFPN)

Whether it is the natural scene or remote sensing scene, FPN is a great strategy for multiscale object detection [49]–[51]. However, the performance of existing FPNs is still poor for extremely imbalanced multiscale and densely distributed objects in RSIs. To address the above problems, the AFPN is proposed to integrate multiscale features sufficiently.

The presented AFPN adopts a similar pipeline as balanced FPN [21] and CE-FPN [20], which has a process of first aggregating multiscale features and then splitting them into feature pyramids. Moreover, we put forward the SRM to perform adaptive balanced fusion between various spatial positions and channels, as shown in Fig. 3. To the best of our knowledge, AFPN can better extract the features of multiscale and dense objects by such a structure.

As shown in Fig. 1, assume that $C_5 \in \mathbb{R}^{C\times H\times W}$; then, we can conclude that $C_4 \in \mathbb{R}^{C\times 2H\times 2W}$, $C_3 \in \mathbb{R}^{C\times 4H\times 4W}$, and $C_2 \in \mathbb{R}^{C\times 8H\times 8W}$. Here, C represents the number of channels

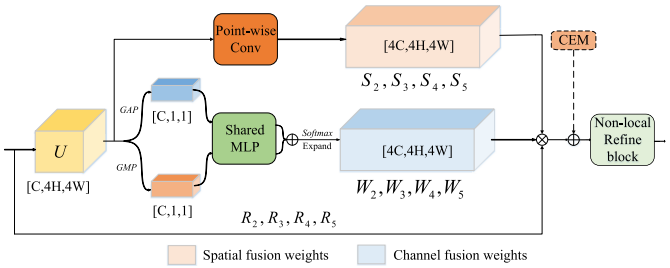

Fig. 3. Illustration of the selective refined module (SRM). It applies two parallel branches to capture spatial fusion weights and channel fusion weights of AFPN.

(equal to 256 in our algorithm), and H and W are 1/32 of the length and the width of the input image, respectively. Specifically, AFPN employs several scale-invariant adaptive pooling layers to generate the same size feature maps $R_2, R_3, R_4, R_5 \in \mathbb{R}^{C\times 4H\times 4W}$ from C_2, C_3, C_4, and C_5, respectively.

After that, AFPN can generate integrated feature map $U \in \mathbb{R}^{C\times 4H\times 4W}$ by elementwise summation from R_2, R_3, R_4, and R_5, i.e.,

$$U = R_2 \oplus R_3 \oplus R_4 \oplus R_5 \tag{6}$$

where $\oplus$ represents elementwise summation.

However, U aggregates semantic information of different scales, which interferes with each other. AFPN adopts SRM to deal with this problem. In short, SRM can learn the fusion weights of spaces and channels via the temporary obtained feature U so as to achieve adaptive fusion.

As for channel fusion weights, SRM includes a GAP and a GMP, as well as a shared MLP to generate channel fusion weights of $4C$ dimensions. The weights W can be calculated as

$$W = \text{softmax}(\text{MLP}(\text{GAP}(U)) + \text{MLP}(\text{GMP}(U))). \tag{7}$$

We split W into four vectors $W_2, W_3, W_4, W_5 \in \mathbb{R}^{C\times 1\times 1}$ and expand them into $\mathbb{R}^{C\times 4H\times 4W}$, which represents the channel weights of R_2, R_3, R_4, and R_5, respectively. By utilizing this operation, AFPN can combine different channel features to facilitate multiscale object detection.

With respect to spatial fusion weights, SRM acquires it conveniently, as illustrated in Fig. 3. SRM adopts a pointwise convolution to capture spatial fusion weights, which efficiently carries out spatial information interaction with low computation. The spatial fusion weights $S \in \mathbb{R}^{4C\times 4H\times 4W}$ are formed as

$$S = \text{softmax}(\text{PointConv}(U)) \tag{8}$$

where PointConv$(\cdot)$ indicates the processing of the 1×1 pointwise convolutional layer with the ReLU function. Besides, SRM applies the softmax function on the same spatial locations of different channels in S. We can also divide S into four vectors $S_2, S_3, S_4, S_5 \in \mathbb{R}^{C\times 4H\times 4W}$, which represents the spatial weights of R_2, R_3, R_4, and R_5. Based on spatial fusion, AFPN discovers more fine-grained spatial features of the objects, which promotes to detect dense objects especially.

With estimated channel fusion weights W_2, W_3, W_4, and W_5 and spatial fusion weights S_2, S_3, S_4, and S_5, the refined integrated feature $\widetilde{U} \in \mathbb{R}^{C\times 4H\times 4W}$ can be generated as

$$\widetilde{U} = \text{NL}\left(\sum_{i=2}^{5} W_i \otimes R_i \otimes S_i + \text{CEM}(F_5)\right) \tag{9}$$

where NL$(\cdot)$ denotes the operation of nonlocal block [52], CEM$(\cdot)$ represents the manipulation of CEM, and F_5 indicates the deepest-level feature map of backbone. The application of NL$(\cdot)$ is inspired by [21], and this refinement step helps AFPN extract more discriminative features and further improve detection results.

AFPN utilizes the SRM block to refine features adaptively. It enables cross-layer correlation between various channels and spatial locations. With the assistance of CEM, AFPN is more capable of achieving selective fusion and separation of multiscale features. The processing of generating the final detection maps of AFPN is described in Section III-C.

C. Context Enhancement Module (CEM)

The deepest-level feature F_5 suffers information loss due to the reduction of channels. Concretely, to unify the detection head, the standard FPN [15] compacts the channels of detected feature maps into C dimensions, where $F_5 \in \mathbb{R}^{8C\times H\times W}$ is compressed into $C_5 \in \mathbb{R}^{C\times H\times W}$. Based on this observation, we propose a novel and effective context module named CEM to address the low efficiency of rich channel information in F_5. It can alleviate the aggregation drawback of insufficient context information of U to improve the detection performance.

CEM boosts the feature representation of F_5 by utilizing different scales of subpixel branches [53] to instill diverse spatial context information into integrated feature U. Theoretically, the spatial context information obtained by subpixel branches can reduce the loss in channels of F_5, thus improving the final feature pyramid simultaneously. Subpixel convolution transforms a tensor with the size of $C \cdot r^2 \times H \times W$ into a one with the size of $C \times H \cdot r \times W \cdot r$, which performs the function of upsampling as

$$F^{\text{SR}}_{x,y,c} = F^{\text{LR}}_{\lfloor x/r\rfloor,\lfloor y/r\rfloor,r\cdot\text{mod}(y,r)+\text{mod}(x,r)+c\cdot r^2} \tag{10}$$

where F^{SR} and F^{LR} indicate the high-resolution feature maps and low-resolution feature maps, respectively.

As shown in Fig. 4, by virtue of applying different proportional subpixel convolutional layers, the multiscale spatial context information is acquired without much computational cost. Then, CEM reconciles the channels of the spatial context information into C dimensions via convolutional layers with the kernel of 1×1. Finally, the spatial context information is unified into a vector with the size of $C \times 4H \times 4W$ by ratio-invariant adaptive pooling. Considering the aliasing effect caused by adaptive pooling, we adopt ASFM rather than simple elementwise summation to combine these contextual features adaptively motivated by AugFPN [18]. The detailed structure of ASFM is shown in Fig. 4. Specifically, ASFM utilizes these contextual features as input and generates several spatial weight maps for them. These weight maps are used to aggregate multibranch contextual features into the adaptive

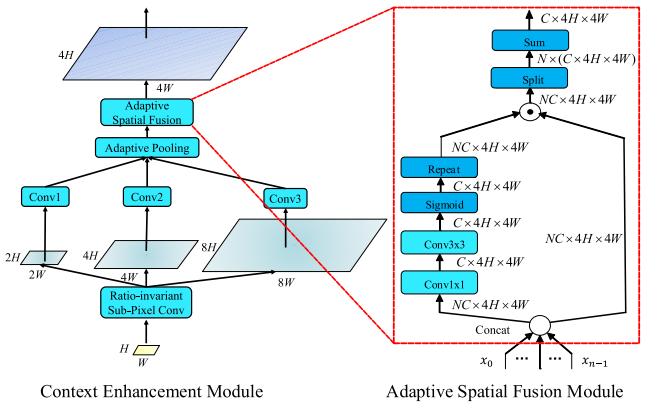

Fig. 4. Left one is the illustration of CEM, and the other is the illustration of ASFM.

feature $\widetilde{U}$ eventually [refer to (9)]. Through such operations, CEM enriches the multiscale semantic information of AFPN, which promotes detection ability for remote sensing objects.

After $\widetilde{U}$ is obtained, the final pyramid $\{P_2, P_3, P_4, P_5\}$ is calculated by multiscale adaptive pooling. Meanwhile, we introduce the skip learning from $\{C_2, C_3, C_4, C_5\}$ to reach $\{P_2, P_3, P_4, P_5\}$, as shown in Fig. 1.

D. Loss Function

In two-stage object detection algorithms, the multitask function is used to balance classification and localization task loss [10]. In this article, the total loss function is defined as

$$L_{\text{total}} = L_{\text{rpn}} + \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{loc}} \tag{11}$$

where L_{rpn}, L_{cls}, and L_{loc} denote the RPN loss, the classification loss, and the regression loss, respectively. In our method, L_{rpn} and L_{cls} are the same as defined in Faster RCNN [10]. Different from [10], we employ balanced smooth L1 loss [21] as L_{loc} since it can accelerate the key regression gradients. Meanwhile, it is capable of rebalancing the samples and achieving balanced training in classification and accurate localization, which is defined as

$$L_{\text{loc}} = \sum_{i \in x,y,w,h} L_b\left(t_i^u - v_i\right) \tag{12}$$

where t^u and v indicate four dimensional coordinate vectors for the predicted b-boxes and the ground truth b-boxes, and $L_b(x)$ is defined as

$$L_b(x) = \begin{cases} \dfrac{\alpha}{b}(b|x|+1)\ln(b|x|+1) - \alpha|x|, & \text{if } |x| < 1 \\ \gamma|x| + C, & \text{otherwise.} \end{cases} \tag{13}$$

Here, the relationship between parameters γ, α, and b satisfy the following equation:

$$\alpha \ln(b+1) = \gamma. \tag{14}$$

We set $\alpha = 0.5$ and $\gamma = 1.5$ in our experiments referring to [21]. For λ_1 and λ_2, we set $\lambda_1 = \lambda_2 = 1$ for simplicity.

IV. Experiments

In this section, we first describe the datasets, evaluation metrics, and so on. Then, all experiments are described and analyzed to verify the performance of the proposed ABNet.

A. Datasets

1) RSOD [40]: It is presented by Wuhan University in 2017, which includes 4993 aircraft, 1586 oil tanks, 191 playgrounds, and 180 overpasses in 2326 RSIs. The image size of this dataset ranges from 512 × 512 to 1961 × 1193 pixels. We adopt the unified strategy in [54] to divide 50% images for training and 50% for testing.

2) NWPU VHR-10 [7]: This dataset, which includes ten categories, is proposed by Cheng *et al.* of Northwestern Polytechnic University, China. The latest version of it contains 1172 images (400 × 400 pixels) cropped from 650 aerial imagery with sizes ranging from 533 × 597 to 1728 × 1028 pixels. According to [48], we split 75% of the dataset (879 images) as the training set and 25% of it (293 images) as the testing set.

3) DIOR [8]: It is the largest dataset for horizontal object detection in geospatial RSIs. It includes 23 463 images (800 × 800 pixels) with 192 472 instances of 20 classes. This dataset is divided into 11 725 images (50% of the dataset) as the training set and the remaining 11 738 images as the testing set.

B. Evaluation Metrics

The widely used performance evaluation metrics for object detection are the average precision (AP) of each class and the mean average precision (mAP) of all classes. For a detector, the mAP is higher, the detection performance is better obviously. The AP and mAP are defined as

$$\text{AP} = \int_0^1 P(R)dR \tag{15}$$

$$\text{mAP} = \frac{1}{N_{\text{cls}}} \sum_{i=1}^{N_{\text{cls}}} \text{AP}_i \tag{16}$$

respectively, where P and R refer to the precision and the recall, and N_{cls} represents the total number of classes. The precision P and the recall R are defined as

$$P = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{17}$$

$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{18}$$

respectively, where N_{TP}, N_{FP}, and N_{FN} denote the number of true positives, false positives, and false negatives, respectively. For a certainly predicted b-box, the conditions for satisfying true positives are given as follows. The first is that the IoU between the predicted b-box and ground truth is not lower than 0.5, and the second is to predict the correct class label.

C. Implementation Details

Our experimental environment is the PyTorch framework (PyTorch1.6) in the Ubuntu 18.04 operating system, and all

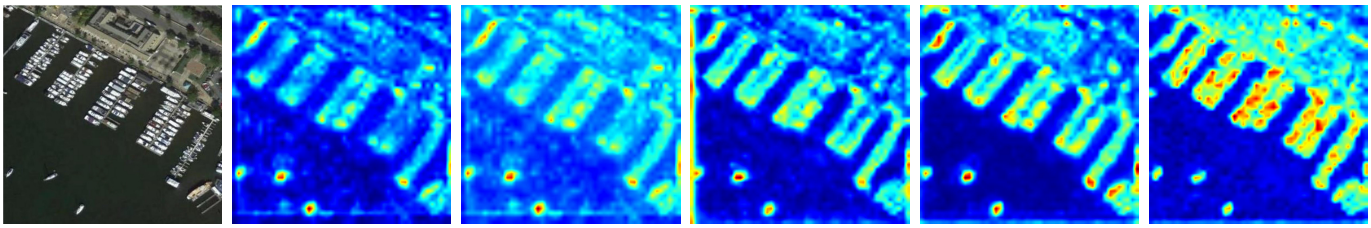

Fig. 5. Features visualizations of different channel attention mechanisms. (a) Input image to network from the NWPU VHR-10 dataset. (b)–(f) Feature maps of the penultimate stage in backbone.

TABLE II
COMPARISON EECA WITH SE [28],CBAM [30], ECA [39] ON NWPU VHR-10 DATASET [7]. THE BEST RESULTS ARE MARKED IN BOLD

Method	mAP_{50} (%)	Params (M)	FLOPs (G)
Baseline(FPN)	90.80	**41.3984**	**134.2953**
+SE	91.26(+0.46)	43.9133	134.3682
+CBAM	91.45(+0.65)	43.9133	134.3707
+ECA	91.55(+0.75)	41.3985	134.3658
+EECA $\gamma = 2$	91.68(+0.88)	41.3986	134.3661
+EECA $\gamma = 1$	**91.92(+1.12)**	41.3987	134.3664
+EECA w/o share	91.88(+1.08)	41.3991	134.3664
+EECA w/o GMP	91.71(+0.91)	41.3987	134.3660

TABLE III
COMPARISON AFPN WITH FPN [15], DFPN [16], PAFPN [17], AND BALANCED FPN [21] ON THE NWPU VHR-10 DATASET [7]. THE BEST RESULTS ARE MARKED IN BOLD

Method	mAP_{50} (%)	Params (M)	FLOPs (G)
Baseline(FPN)	90.80	**41.3984**	**134.2953**
DFPN	91.73(+0.93)	44.6808	161.9187
PAFPN	91.78(+0.98)	45.5290	173.3881
Balanced FPN	91.96(+1.16)	41.5301	134.6246
AFPN(Ours)	**92.37+(1.57)**	42.4483	138.9104

experiments are performed on two NVIDIA TITAN RTX GPUs with 24-GB memory per GPU. ResNet50 [43] pretrained on ImageNet-1K classification task is the backbone. We adopt Kaiming normal to initialize new layers and the stochastic gradient descent (SGD) algorithm to optimize the parameters of the model. The initial learning rate of SGD is 0.02. The weight decay and momentum of SGD are 0.0001 and 0.9, respectively. In all experiments, we employ 0.5 horizontal flips as data augmentation without other tricks. The number of total epochs is 20, and at the eighth and 14th epochs, the learning rate is reduced to 0.1 and 0.01 times, respectively. The size of input images for the network is 800×800 for three datasets. The batch size per GPU is 4 for DIOR and RSOD, while it is 12 for NWPU VHR-10. All other hyperparameters are consistent with the standard Faster RCNN [10] with FPN [15].

D. Model Analysis

1) Comparison of EECA: We select several state-of-the-art channel attention mechanisms (i.e., SE [28], CBAM [30], and ECA [39]) for fair comparison with EECA on NWPU VHR-10. Experimental results are illustrated in Table II, where "EECA $\gamma = 1$" indicates our proposed EECA, and "EECA $\gamma = 2$" is intended for a fair comparison with ECA as it calculates kernel of 1-D convolution with $\gamma = 2$. In addition, "EECA w/o share" denotes that the GAP and GMP adopt two separate parallel paths instead of shared 1-D convolutional layers. "EECA w/o GMP" is defined as only using GAP to obtain channel information without GMP. It is obvious that EECA outperforms other channel attention mechanisms. In particular, the proposed "EECA $\gamma = 1$" achieves the best experimental results (1.12% mAP ↑). Meanwhile, "EECA $\gamma = 1$" yields more performance gain than "EECA w/o share," which shows the effectiveness of shared 1-D convolutional layers. "EECA $\gamma = 1$" has a better performance than "EECA w/o GMP," which shows that GMP is, indeed, able to take the most significant knowledge of each channel into account. Through the comparison of parameters and FLOPs, we find that the proposed EECA is more lightweight with better performance than SE and CBAM. The reason for the poor performance of SE and CBAM is that they introduce a large number of additional parameters, which makes them difficult to optimize remote sensing tasks. In comparison with ECA, our EECA introduces two 1-D convolutional layers with stronger nonlinear expression ability to capture larger local cross-channel interaction, thus producing better detection results for RSIs. As illustrated in Fig. 5, EECA is able to distinguish foreground and background more significantly and generates higher activation response values for dense remote sensing objects than SE [28], CBAM [30], and ECA [39]. We conclude that the nonlinear modeling between local cross-channel in CNNs is very crucial for RSI processing. It actually mitigates the negative impact of background clutter and strengthens the feature extraction of CNNs for RSIs.

2) Comparison of AFPN: We compare AFPN against several state-of-the-art pyramid networks, such as DFPN [16], PAFPN [17], and Balanced FPN [21] on NWPU VHR-10, to reveal the superiority of AFPN. The experimental results are illustrated in Table III. AFPN increases performance by 1.57% mAP, which is better than other FPNs. Most importantly, AFPN reduces the model parameters and FLOPs than DFPN [16] and PAFPN [17], while bringing about only a few parameters and FLOPs compared with the Balanced FPN [21]. In fact, the prime key of AFPN is that it not

TABLE IV

ABLATION STUDY ON THE NWPU VHR-10 DATASET [7] AND THE RSOD DATASET [40]. THE BEST RESULTS ARE MARKED IN BOLD

EECA	AFPN	CEM	mAP on NWPU VHR-10 (%)	mAP on RSOD (%)	Parameters (M)	FLOPs (G)
×	×	×	90.80	90.91	**41.398**	**134.295**
✓	×	×	91.92	91.90	41.399	134.366
×	✓	×	92.37	92.29	42.448	138.910
×	×	✓	91.83	92.14	41.775	135.433
✓	✓	×	93.41	93.44	42.448	138.981
×	✓	✓	93.24	93.37	42.824	140.057
✓	✓	✓	**94.21**	**94.17**	42.825	141.374

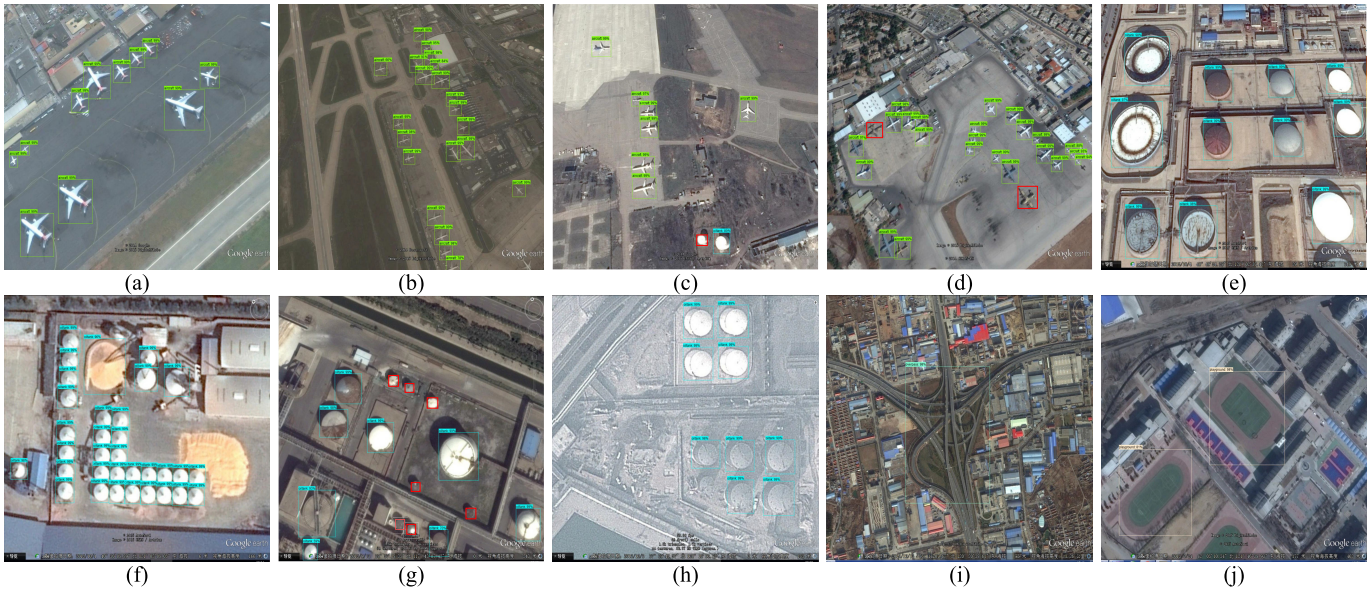

Fig. 6. Some representative detection results of our ABNet on the RSOD dataset. These sample images are uniform in size for esthetic layout. (a) and (b) Aircraft. (c) Aircraft and oil tank. (d) Aircraft. (e)–(h) Oil tank. (i) Overpass. (j) Playground. Red boxes are the missing predictions.

only combines features of different stages via adaptive pooling but also performs spatial and channel fusion via SRM block. Specifically, SRM solves the features aliasing problem of spatial positions and channels between multiscale feature maps of RSIs. The comparative FPNs [16], [17], [21] cannot address this problem well, which results in less significant performance improvements than our AFPN.

E. Ablation Study

We set up ablation experiments on RSOD and NWPU VHR-10 datasets to prove the effects of EECA, AFPN, and CEM in our algorithm. Table IV displays the results of ablation experiments.

1) Baseline Setup: Our baseline is Faster RCNN [10] with FPN [15]. It uses ResNet50 as the backbone, multiscale RoI Align for regional feature extraction, and balanced smooth L1 loss [21] as regression task loss. In all experiments, all hyperparameters are consistent for a fair comparison. Baseline reaches 90.80% mAP and 90.91% mAP on NWPU VHR-10 and RSOD, respectively.

2) Effect of EECA: The proposed EECA in ABNet is designed to restrain the interference of RSIs with complicated backgrounds. In Table IV, our method (only with EECA) can reach 91.92% mAP (1.12%↑) and 91.90% mAP (0.99%↑) on the NWPU VHR-10 dataset and the RSOD dataset, respectively. This validates that EECA is an efficient channel attention mechanism with only adding 324 parameters. It really helps ResNet [43] extract more fine-grained features from input RSIs. When our method merges EECA and AFPN together, it achieves 93.41% mAP and 93.44% mAP on NWPU VHR-10 and RSOD, respectively, which indicates that the combination of EECA and AFPN can achieve better detection.

TABLE V

COMPARISON WITH STATE OF THE ARTS ON RSOD [40]. * DENOTES OUR IMPLEMENTATION, AND THE BEST RESULTS ARE MARKED IN BOLD

Method	Aircraft	Oil tank	Overpass	Playground	mAP
Faster RCNN* [10]	71.30	90.70	90.90	99.70	88.10
Sig-NMS [25]	80.60	90.60	87.40	99.10	89.40
YOLOv3* [13]	88.60	94.50	75.90	**99.90**	89.70
FPN* [15]	90.58	94.47	80.18	98.49	90.91
RFN [27]	79.10	90.50	**100**	99.70	92.30
SSAFNet [54]	**95.75**	98.39	84.66	92.50	92.82
CF2PN [45]	95.52	**99.42**	83.82	95.68	93.61
ABNet* (Ours)	91.49	96.14	89.61	99.44	**94.17**

TABLE VI

COMPARISON WITH STATE OF THE ARTS ON NWPU VHR-10 [7]. * DENOTES OUR IMPLEMENTATION, AND THE BEST RESULTS ARE MARKED IN BOLD

Method	Backbone	Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle	mAP
RICNN [55]	AlexNet	88.71	78.34	86.33	89.09	42.33	56.85	87.72	67.47	62.31	72.01	73.11
RICAOD [48]	ZFNet	99.70	90.80	90.61	92.91	90.29	80.13	90.81	80.29	68.53	87.14	87.12
YOLOv3* [13]	DarkNet53	99.55	81.82	80.30	98.26	80.56	81.82	99.47	74.31	**89.61**	86.98	87.27
Faster RCNN* [10]	ResNet50	**100**	85.28	**100**	95.93	87.59	92.08	99.73	92.11	43.37	86.60	88.30
FMSSD [34]	VGG16	99.70	89.90	90.30	98.20	86.00	**96.80**	99.60	75.60	80.10	88.20	90.40
FPN* [15]	ResNet50	**100**	90.86	99.99	96.84	90.67	95.05	**100**	93.67	50.86	90.19	90.80
CAD-Net [32]	ResNet101	97.00	77.90	95.60	93.60	87.60	87.10	99.60	**100**	86.20	89.90	91.50
SCRDet [35]	ResNet101	**100**	89.40	97.20	97.00	83.20	87.50	99.20	99.40	74.50	90.10	91.75
GA-RetinaNet [56]	ResNet101	99.99	84.28	97.92	96.53	96.98	85.12	95.34	89.72	81.32	91.85	91.91
FCOS [57]	ResNet101	99.99	85.21	96.94	97.75	95.80	80.34	99.67	95.04	81.82	88.92	92.14
auto-MSNet [31]	DarkNet53	99.00	85.30	93.30	**99.50**	95.10	94.60	98.80	86.90	85.20	86.80	92.50
CANet [37]	ResNet101	99.99	85.99	99.27	97.28	97.80	84.77	98.38	90.38	89.16	90.25	93.33
ABNet* (Ours)	ResNet50	**100**	**92.58**	97.77	97.76	**99.26**	95.98	99.86	94.26	69.04	**95.62**	**94.21**

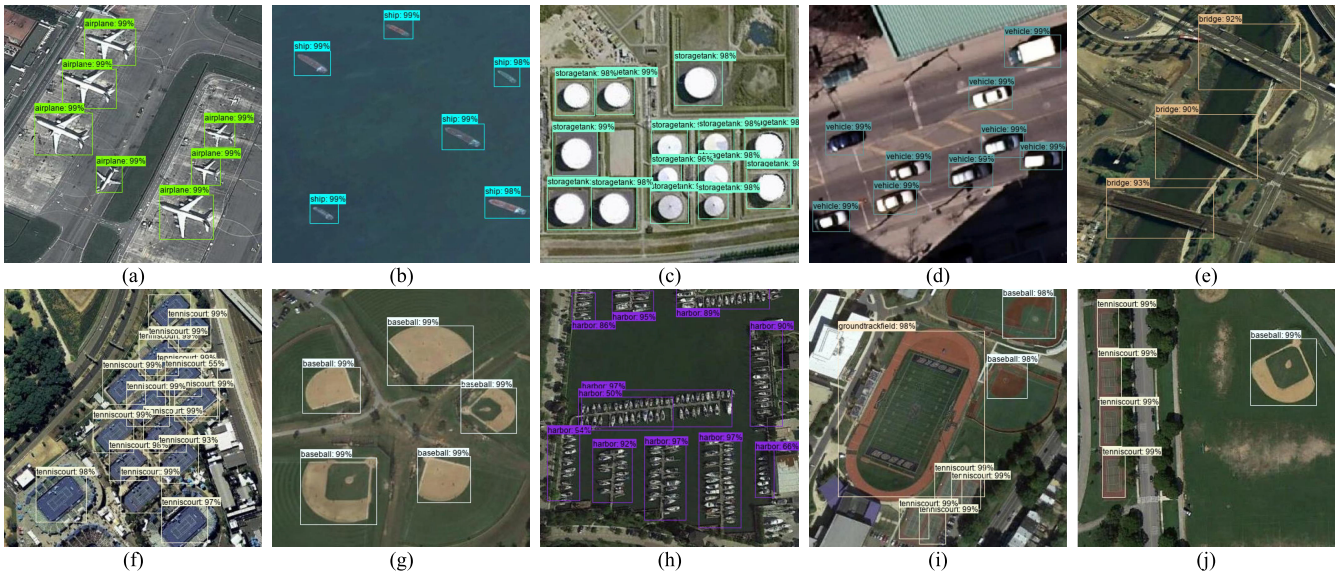

Fig. 7. Some representative detection results of our ABNet on the NWPU VHR-10 dataset. (a) Airplane. (b) Ship. (c) Storage tank. (d) Vehicle. (e) Bridge. (f) Tennis court. (g) Baseball diamond. (h) Harbor. (i) Ground track field, tennis court, and baseball diamond. (j) Tennis court and baseball diamond.

3) Effect of AFPN: To detect multiscale and dense objects efficiently in RSIs, AFPN is proposed. We compare it against the baseline to validate the strategy of AFPN. Our method only with AFPN can obtain 92.37% mAP (1.57%↑) on the NWPU VHR-10 dataset and 92.29% mAP (1.38%↑) on the RSOD dataset, respectively. It illustrates that the simple strategy of spatial localities and channels of multiscale feature maps via AFPN really contributes to the network for interpreting RSIs. In addition, AFPN only adds 1.05M parameters approximately, which is less than 2.5% of the baseline.

4) Effect of CEM: Considering that the reduction of channels in the original FPN would lose vital semantic information, we design the CEM to instill diverse spatial context information into AFPN. To validate this point, we conduct experiments involving only CEM. When adding CEM to the baseline to enhance C_3, the results show the improvement of 1.03% mAP on NWPU VHR-10 and 1.23% mAP on RSOD with increasing only 0.377M parameters and 1.138G FLOPs. The employment of CEM enables the network to pay more attention to deep semantic information. In fact, CEM further improves the detection performance of AFPN. When integrating both AFPN and CEM into the baseline, we observe that the detection performance is further improved, reaching 93.24% and 93.37% mAP on NWPU VHR-10 and

TABLE VII

COMPUTATION TIME COMPARISON OF EIGHT APPROACHES ON THE NWPU VHR-10 DATASET [7]

Method	Average computation time per image (seconds)
RICNN [55]	8.47
RICAOD [48]	2.89
GA-RetinaNet [56]	0.18
auto-MSNet [31]	0.14
CANet [37]	0.10
FCOS [57]	0.09
ABNet (Ours)	0.07
YOLOv3 [13]	0.05

TABLE VIII

COMPARISON WITH STATE OF THE ARTS ON DIOR [8]. * DENOTES OUR IMPLEMENTATION, AND THE BEST RESULTS ARE MARKED IN BOLD

Method	AL	AT	BF	BC	BG	CM	DM	EA	ES	GC	GF	HB	OP	SP	SD	ST	TC	TS	VH	WM	mAP
RICAOD [48]	42.2	69.7	62.0	79.0	27.7	68.9	50.1	60.5	49.3	64.4	65.3	42.3	46.8	11.7	53.5	24.5	70.3	53.3	20.4	56.2	50.9
YOLOv3* [13]	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	**87.4**	70.6	68.7	87.3	29.4	48.3	78.7	57.1
FPN* [15]	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2	65.1
Eff-Det [19]	86.5	57.4	75.7	85.2	33.5	75.4	65.6	80.1	67.4	58.3	71.4	35.6	50.6	78.8	**90.3**	61.8	82.9	54.6	30.0	81.5	66.1
CF2PN [45]	78.3	78.3	76.5	88.4	37.0	71.0	59.9	71.2	51.2	75.6	77.1	56.8	58.7	76.1	70.6	55.5	88.8	50.8	36.9	86.4	67.3
O^2-DNet [24]	61.2	80.1	73.7	81.4	45.2	75.8	64.8	81.2	**76.5**	79.5	79.7	47.2	59.3	72.6	70.5	53.7	82.6	55.9	49.1	77.8	68.4
PANet* [17]	62.4	76.3	71.6	87.3	48.6	79.3	65.5	75.9	72.8	76.4	**82.5**	47.2	60.6	72.0	68.7	62.6	81.2	56.3	50.5	88.0	69.3
FCOS [57]	61.1	82.6	76.6	87.6	42.8	**80.6**	64.1	79.1	67.2	**82.0**	79.6	46.4	57.8	72.1	64.8	63.4	85.2	62.8	43.8	87.5	69.4
SB-MSN [38]	79.6	82.2	76.4	89.8	45.6	78.2	64.8	58.9	59.3	79.2	82.4	51.8	60.8	74.4	79.7	66.4	85.6	**65.4**	45.1	79.9	70.3
LRCNN* [21]	60.8	79.8	71.7	87.8	49.5	79.8	64.8	82.0	74.8	79.7	**82.5**	42.9	**63.0**	72.0	75.4	62.7	81.3	64.3	49.9	88.5	70.7
ASSD [46]	85.6	82.4	75.8	89.5	40.7	77.6	64.7	67.1	61.7	80.8	78.6	**62.0**	58.0	84.9	76.7	65.3	87.9	62.4	44.5	76.3	71.1
FSoD-Net [44]	**88.9**	66.9	**86.8**	**90.2**	45.5	79.6	48.2	**86.9**	75.5	67.0	77.3	53.6	59.7	78.3	69.9	**75.0**	**91.4**	52.3	52.0	**90.6**	71.8
ABNet* (Ours)	66.8	**84.0**	74.9	87.7	**50.3**	78.2	**67.8**	85.9	74.2	79.7	81.2	55.4	61.6	75.1	74.0	66.7	87.0	62.2	**53.6**	89.1	**72.8**

"Eff-Det" means EfficientDet, "LRCNN" means Libra RCNN. AL: Airplane. AT: Airport. BF: Baseball Field. BC: Basketball Court. BG: Bridge. CM: Chimney. DM: Dam. EA: Expressway Service Area. ES: Expressway toll Station. GC: Golf Course. GF: Ground Track Field. HB: Harbor. OP: Overpass. SP: Ship. SD: Stadium. ST: Storage Tank. TC: Tennis Court. TS: Train Station. VH: Vehicle. WM: Wind Mill.

Fig. 8. Some representative detection results of ABNet on the DIOR dataset. (a) Airplane and Vehicle. (b) Bridge, stadium, ground track field, and overpass. (c) Baseball field, tennis court, and basketball court. (d) Bridge, airport, and expressway toll station. (e) Windmill. (f) Vehicle, basketball court, and tennis court. (g) Ground track field and dam. (h) Ship and harbor. (i) Chimney and storage tank. (j) Golf field. (k) Storage tank. (l) Train station. Red boxes are the missing predictions.

TABLE IX

AP_s, AP_m, AP_l, AR_s, AR_m, AR_l, AND mAP_{50} OF DIFFERENT METHODS ON THE DIOR DATASET [8]. THE BEST RESULTS ARE MARKED IN BOLD

Method	$AP_s(\%)$	$AP_m(\%)$	$AP_l(\%)$	$AR_s(\%)$	$AR_m(\%)$	$AR_l(\%)$	$mAP_{50}(\%)$
YOLOv3 [13]	8.2	26.7	45.9	15.1	38.5	59.0	57.13
FPN [15]	13.2	36.0	57.7	**21.0**	45.7	66.5	65.10
PANet [17]	**13.5**	36.6	59.2	20.2	45.7	67.4	69.30
ASSD [46]	10.5	**37.7**	63.8	14.9	**47.2**	70.4	71.13
ABNet (Ours)	13.3	37.6	**66.0**	20.5	47.1	**73.1**	**72.77**

RSOD, respectively. The reason is that CEM helps AFPN catch multiscale semantic information from the deepest-level layer of the backbone, which is remarkable for multiscale object detection in RSIs.

F. Comparison With State-of-the-Art Methods

The comparison experiments of ABNet and other advanced detectors on RSOD [40], NWPU VHR-10 [7], and DIOR [8] are analyzed in this subsection. The results are shown in Tables V–VIII.

1) Results on RSOD: As displayed in Table V, ABNet reaches 94.17% mAP in the RSOD dataset, which achieves the best performance among competitors. Although our algorithm does not offer the highest AP in any category, the result is balanced with AP above 90% in almost all categories. By contrast with FPN (baseline), the ABNet performs better in all categories, and it upgrades the AP by percentages of 0.91, 1.67, 9.43, and 0.95, respectively, which demonstrates the effectiveness of our proposed modules. Specifically, the detection performance of overpass is directly boosted by 9.43%, which is mainly derived from the promotion of the EECA mechanism. EECA suppresses negative information of complex backgrounds by capturing local cross-channel correlation so that ABNet offers a great performance improvement on large-scale categories.

2) Results on NWPU VHR-10: In Table VI, we compare our ABNet with some state of the arts, including one-stage detectors: FMSSD [34], YOLOv3 [13], GA-RetinaNet [56], auto-MSNet [31], CANet [37] and two-stage detectors: RICNN [55] and RICAOD [48], Faster RCNN [10], FPN [15], SCRDet [35], and CAD-Net [32]. In addition, we add the state-of-the-art anchor-free algorithm FCOS [57] to the comparison. Among all competitors, ABNet achieves the highest AP in the densely distributed ships (92.58%) and vehicles (95.62%), which demonstrates its effectiveness for dense object detection. Unfortunately, CANet [37] describes the centerness of symmetrical objects for RSIs while performing unsatisfactory results in densely arranged objects (e.g., the AP in ships is 6.59% lower than ours, and the AP in vehicles is 5.37% lower than ours). In addition, ABNet produces a significant performance improvement in the large-scale bridge category with an 18.18% increase in AP compared with FPN (baseline), which also shows the efficiency of the EECA mechanism. Furthermore, the overall detection performance is 5.91% higher than Faster RCNN [10] and 3.41% higher than FPN [15]. It should be noted here that our approach adopts the standard anchor definition according to Faster RCNN [10]. For objects with large aspect ratios, such as bridges, our baseline encounters a bottleneck, which only achieves the detection performance of 50.86% AP. The above results reflect that our ABNet achieves effective multiscale object detection by constructing various submodules and combining them with appropriate loss functions.

Table VII shows the average computation time of eight approaches on the NWPU VHR-10 dataset, which is widely adopted in RSOD [55]. As a fast one-stage detector, YOLOv3 [13] has the lowest running time among competitors. The computation time of ABNet has only a 0.02 s gap with YOLOv3 and outperforms the other six algorithms. It is worth mentioning that our ABNet as a two-stage algorithm is faster than single-stage methods: FCOS [37] and CANet [57]. ABNet has such excellent performance in keeping speed/accuracy tradeoff because it uses ResNet50 as backbone instead of heavy ResNet101, and the proposed structures (EECA, AFPN, and CEM) have low parameter costs.

3) Results on DIOR: We evaluate the ABNet on DIOR and compare it against the latest approaches, encompassing RICAOD [48], YOLOv3 [13], EfficientDet [19], FPN [15], CF2PN [45], O^2-DNet [24], SB-MSN [38], PANet [17], Libra RCNN [21], ASSD [46], FCOS [57], and FSoD-Net [44]. ABNet is the only method that exceeds 72% mAP, as shown in Table VIII, and it achieves the best results in four of the total 20 categories, i.e., airport, bridge, dam, and vehicle. Although FSoD-Net achieves the highest detection accuracy in many classes, it is still 1% lower than ABNet in overall mAP. The main reason is that FSoD-Net [44] does not solve the sophisticated background problem well, which leads to low performance in large-scale objects (e.g., 66.9% AP in airports, 45.5% AP in bridges, and 48.2% AP in dams). ABNet increases detection accuracies by 3.3% and 10.5% on dense categories of ship and vehicle compared with FPN (baseline), which certifies the effectiveness of ABNet for dense object detection. This is primarily because AFPN makes our detector focus on more spatial characteristics of clustered objects; hence, considerable detection performance is accomplished. Meanwhile, our ABNet obtains the best detection accuracies for large-scale objects among competitors, i.e., 84.0% mAP for airports and 67.8% mAP for dams. The reasonable explanation is that its EECA mechanism highlights the large objects in RSIs and suppresses negative information of complicated background. We notice that ABNet performs lower than the most advanced comparison algorithms on small objects. For example, it reaches 66.8% in airplanes (worse than 88.9% of FSoD-Net) and 75.1% in ships (worse than 87.4% of

YOLOv3). It may be because AFPN tends to take into account the characteristics of large-scale objects when performing feature selection, which contributes less semantic information of small objects.

For further assessment, we adopt AP_s, AP_m, AP_l, AR_s, AR_m, and AR_l in COCO evaluation criteria to quantitatively report the performance of several state-of-the-art algorithms. These comparison approaches include YOLOv3 [13], FPN [15], PANet [17], and ASSD [46]. The details are illustrated in Table IX. We employ the PyTorch code[2] to implement YOLOv3 [13] and adopt MMDetection framework[3] to implement PANet [17]. Besides, the authors of ASSD [46] provide us with relevant data. By comparison, we find that ABNet achieves the best performance on AP_l and AR_l. However, it performs a little worse on AP_s, AP_m, AR_s, and AR_m in comparison to others. This fact reveals that ABNet has room for improvement in the detection of small and middle-sized objects.

G. Qualitative Analysis

As shown in Figs. 6–8, ABNet achieves excellent detection performance on NWPU VHR-10. For the RSOD, ABNet can detect a variety of objects with different scales and shapes robustly. For example, the aircraft in Fig. 6(a) and (b) with various scales distribution can be well detected. Unfortunately, several small objects are missed in Fig. 6(c), (d), and (g). As shown in Fig. 8, ABNet can overcome the disturbance of complex background, multiscale, and dense object distribution, which achieves decent qualitative detection performance for the most challenging dataset DIOR. Specifically, ABNet is able to detect extremely large "golf field," "train station," and densely packed "vehicle," "ship," "airplane," and "storage tank." However, ABNet fails to detect tiny objects, as shown in Fig. 8(a) and (j).

V. Conclusion

In this article, an improved detector ABNet with three upgrades based on Faster RCNN is proposed for RSIs. First, to explore correlations between local cross-channels, the EECA mechanism is designed to achieve more effective channel feature extraction capability for ResNet. EECA mechanism highlights the large objects in RSIs and suppresses negative information of complicated background. Second, AFPN is developed, which only introduces an MLP, a pointwise convolution, and a nonlocal block to integrate feature maps of various scales efficiently. Third, CEM is deployed to combine the deepest-level features of the backbone into AFPN and coalesce sufficient contextual information. Experiments on three public benchmarks prove that ABNet significantly outperforms many state-of-the-art algorithms. Our method only introduces less than 1.5M extra parameters than baseline, which maintains a decent running speed. We find that the detection performance of ABNet for small objects is not significantly improved. Therefore, how to design a lightweight and better detector for small objects will be further investigated in our future work. In addition, we will explore the performance of the EECA mechanism in other remote sensing tasks.

[2]https://github.com/ultralytics/yolov3

[3]https://github.com/open-mmlab/mmdetection

References

[1] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 230–241, Jan. 2017.

[2] W. Xie, J. Lei, S. Fang, Y. Li, X. Jia, and M. Li, "Dual feature extraction network for hyperspectral image analysis," *Pattern Recognit.*, vol. 118, Apr. 2021, Art. no. 107992.

[3] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, "Hyperspectral pansharpening with deep priors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1529–1543, May 2020.

[4] G. Ganci, A. Cappello, G. Bilotta, and C. Del Negro, "How the variety of satellite remote sensing data over volcanoes can assist hazard monitoring efforts: The 2011 eruption of nabro volcano," *Remote Sens. Environ.*, vol. 236, Jan. 2020, Art. no. 111426.

[5] W. Xie, X. Zhang, Y. Li, J. Lei, J. Li, and Q. Du, "Weakly supervised low-rank representation for hyperspectral anomaly detection," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 3889–3900, Aug. 2021.

[6] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1457–1470, May 2018.

[7] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[8] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[9] J. Ding *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 6, 2021, doi: 10.1109/TPAMI.2021.3117983.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "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, Jun. 2017.

[11] 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.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

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

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[15] T.-Y. Lin, P. Dollar, 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.

[16] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, "Deep feature pyramid reconfiguration for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 169–185.

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

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

[19] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[20] Y. Luo *et al.*, "CE-FPN: Enhancing channel information for object detection," 2021, *arXiv:2103.10643*.

[21] J. Pang *et al.*, "Towards balanced learning for instance recognition," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1376–1393, May 2021.

[22] J. Chen, B. Luo, Q. Wu, J. Chen, and X. Peng, "Overlap sampler for region-based object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 767–775.

[23] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, early access, Apr. 20, 2021, doi: 10.1109/TMM.2021.3074273.

[24] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

[25] R. Dong, D. Xu, J. Zhao, L. Jiao, and J. An, "Sig-NMS-Based faster R-CNN combining transfer learning for small target detection in VHR optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8534–8545, Nov. 2019.

[26] H. Guo, X. Yang, N. Wang, B. Song, and X. Gao, "A rotational libra R-CNN method for ship detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5772–5781, Aug. 2020.

[27] K. Zhou, Z. Zhang, C. Gao, and J. Liu, "Rotated feature network for multiorientation object detection of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 33–37, Jan. 2021.

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

[29] J. Wang, Y. Wang, Y. Wu, K. Zhang, and Q. Wang, "FRPNet: A feature-reflowing pyramid network for object detection of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, early access, Dec. 8, 2021, doi: 10.1109/LGRS.2020.3040308.

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

[31] S. Zhang, X. Mu, G. Kou, and J. Zhao, "Object detection based on efficient multiscale auto-inference in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1650–1654, Sep. 2021.

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

[33] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to local: Clip-LSTM-based object detection from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Mar. 25, 2021, doi: 10.1109/TGRS.2021.3064840.

[34] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, Dec. 2020.

[35] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8231–8240.

[36] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8311–8320.

[37] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "CANet: Centerness-aware network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Apr. 2, 2021, doi: 10.1109/TGRS.2021.3068970.

[38] W. Han *et al.*, "Improving training instance quality in aerial image object detection with a sampling-balance-based multistage network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10575–10589, Dec. 2021.

[39] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[40] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representat. (ICLR)*, 2015, pp. 1–15.

[42] S. Li, Q. Yan, and P. Liu, "An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism," *IEEE Trans. Image Process.*, vol. 29, pp. 8467–8475, 2020.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[44] G. Wang *et al.*, "FSoD-Net: Full-scale object detection from optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022, doi: 10.1109/TGRS.2021.3064599.

[45] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, "CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection," *Remote Sens.*, vol. 13, no. 5, p. 847, Feb. 2021.

[46] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, early access, Jun. 29, 2021, doi: 10.1109/TGRS.2021.3089170.

[47] X. Zeng *et al.*, "Crafting GBD-Net for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2109–2123, Sep. 2017.

[48] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[49] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 22, 2021. 10.1109/TPAMI.2021.3053577.

[50] J. Cao, Y. Pang, J. Han, and X. Li, "Hierarchical shot detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9704–9713.

[51] Y. Miao, Z. Lin, X. Ma, G. Ding, and J. Han, "Learning transformation-invariant local descriptors with low-coupling binary codes," *IEEE Trans. Image Process.*, vol. 30, pp. 7554–7566, 2021.

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

[53] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[54] Y. Guo, J. Ji, X. Lu, H. Xie, and X. Tong, "Geospatial object detection with single shot anchor-free network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Oct. 2020, pp. 280–283.

[55] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[56] J. Wang, K. Chen, S. Yang, C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2960–2969.

[57] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

Yanfeng Liu (Student Member, IEEE) received the B.E. degree in computer science and technology from Northeast Forestry University, Harbin, China, in 2021. He is pursuing the M.S. degree in computer science and technology with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

His research interests include computer vision, pattern recognition, and remote sensing.

Qiang Li (Graduate Student Member, IEEE) received the B.E. degree in measurement and control technology and instrument from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2015, and the M.S. degree in communication and transportation engineering from Chang'an University, Xi'an, in 2018. He is pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an.

His research interests include hyperspectral image processing and computer vision.

Yuan Yuan (Senior Member, IEEE) is a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored over 150 papers, including over 100 in reputable journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her research interests include visual information processing and image/video content analysis.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore County, Baltimore, MD, USA, in 2000.

She is a Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is a fellow of SPIE—International Society for Optics and Photonics and a member of the IEEE Periodicals Review and Advisory Committee. She was a recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS). She served as the Co-Chair for the Data Fusion Technical Committee of the IEEE GRSS from 2009 to 2013. She was the Chair with the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She was the General Chair for the fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing held in Shanghai, China, in 2012. She served as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS), the *Journal of Applied Remote Sensing*, and the IEEE SIGNAL PROCESSING LETTERS. From 2016 to 2020, she was the Editor-in-Chief of the IEEE JSTARS.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.