ABNet: Adaptive Balanced Network for Multi-scale Object Detection in Remote Sensing Imagery

Yanfeng Liu, *Student Member, IEEE,* Qiang Li, *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 is a challenging task mainly due to: 1) complicated background of remote sensing images; 2) extremely imbalanced scale and sparsity distribution of remote sensing objects. Existing methods can not effectively solve these problems with excellent detection accuracy and rapid speed. To address these issues, we propose an Adaptive Balanced Network in this paper. Firstly, we design an Enhanced Effective Channel Attention (EECA) mechanism to improve the feature representation ability of backbone, which can alleviate the obstacles of complex background on foreground objects. Then, to combine multi-scale 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 multi-scale 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*—Remote sensing image, multi-scale object detection, local cross-channel attention, adaptive feature pyramid, context exploitation.**

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 multi-objects, which has been widely used in hazard response [4], urban monitoring [5], traffic control [6], etc. Although many algorithms have been proposed for RSOD, especially for large-scale RSIs, this task still remains challenges mainly due to complex scenes and multi-scale objects.

Different from natural scene images, RSIs are commonly captured from satellites with wide view, 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 main obstacles for object detection in RSIs, which make 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 Region Proposal Network (RPN) to gather proposals instead of sliding window. The typical one-stage methods mainly include YOLO [11]–[13], RetinaNet [14], etc. 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 multi-scale objects. For instance, Faster RCNN [10], 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 with multi-scale features for detection. A number of improved FPNs have been widely studied [18]–[20] thereafter. Nevertheless, FPNs only address multi-scale 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 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 field.

Considering the various characteristics of RSIs, plenty of improved algorithms (based on natural scene object detection methods) have been applied to remote sensing area, such as [24]–[26], etc. To alleviate the object confusion caused by complex background in RSIs, RFN [27] embeds Squeeze and Excitation (SE) blocks [28] into detector for features selection.

This work was supported by the National Natural Science Foundation of China under Grants U21B2041, U1864204, 61632018, and 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, P. R. China (e-mail: liuyanfeng99@gmail.com, liqmges@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).

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

Wang *et al.* propose FRPNet [29] that adds 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 multi-scale 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 with 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 existing public remote sensing datasets provide ground truth for clusters.

Nonetheless, most of these approaches cannot perform very well in sophisticated scenes, especially for multi-scale 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 with these problems for RSOD still need more research efforts.

To address the above-mentioned issues, we propose an Adaptive Balanced Network (ABNet) with a series of components to improve the detection accuracy and maintain superior running speed. Initially, to relieve 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 multi-scale feature maps by adaptive pooling. After that, we present a novel Selective Refined Module (SRM) to reconstruct AFPN. Furthermore, a context module (CEM) is designed to mitigate the lack of contextual information in integrated features and construct multi-scale pyramid network for detection. The multi-level RPN [15] is utilized to create and identify proposal candidates according to the multi-scale 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 paper are summarized as follows.

- Aiming at large-scale RSIs with complex background, we design EECA mechanism to extract fine-grained features. It is lighter to deploy in deep CNNs with bringing few parameters as well as achieving better performance than SE [28], CBAM [30] and ECA [39].
- To solve multi-scale and dense object detection efficiently, a novel pyramid network (AFPN) is presented. It contains Selective Refined Module (SRM) to refine different feature maps selectively.
- We propose the Context Enhancement Module (CEM) to address the low efficiency of rich channel information of 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 NWPU VHR-10 [7] dataset and RSOD [40] dataset, respectively. Meanwhile, it achieves 72.8% mAP on 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

Attention module in deep learning is first proposed in 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 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 non-linear 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, 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.
3) ECANet [39] deploys 1D convolutional layer with 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.

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

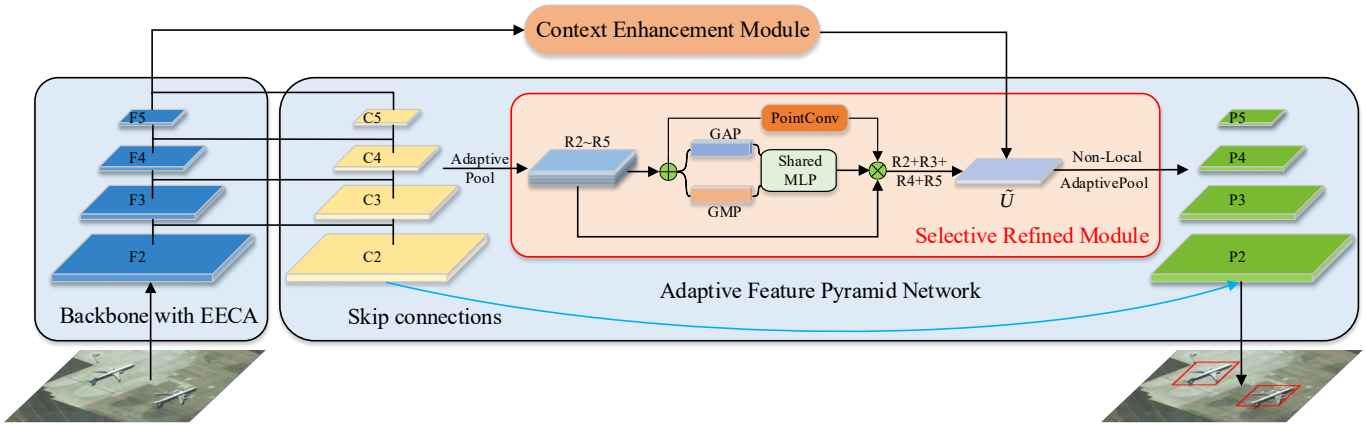


Fig. 1. Illustration of our proposed ABNet. For an input remote sensing image, ABNet first applies ResNet50 backbone (modified by EECA) to extract multi-scale feature maps $\{F_2, F_3, F_4, F_5\}$. Then, Adaptive Feature Pyramid Network utilizes its selective fusion strategy to produce the aggregated feature map $\widetilde{U}$. After that, ABNet uses the Context Enhanced Module to ameliorate $\widetilde{U}$ and obtain multi-scale features $\{P_2, P_3, P_4, P_5\}$ for detection.

Notoriously, the above mechanisms are all designed for natural images. However, they result in sub-optimal performance in RSIs due to the inconsistent data distribution. Besides, the huge amount of extra parameters (illustrated in Table I, e.g., SE, CBAM) increase the difficulty of fine-tuning the networks in aerial tasks. In the next section, 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

Feature Pyramid Network (FPN) is first presented in [15]. It is an elegant method to solve multi-scale object detection by using up-sampling and element-wise 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 global attention model (SE blocks [28]) and local reconfiguration model (residual blocks), which can interact with features across locations and scales. PAFPN [17] adds a bottom-up path to better integrate multi-scale 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 features 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 multi-scale and dense objects efficiently.

C. Context Exploitation

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 background. 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 take into account the reduction of channels in deep-level features of backbone. In this paper, we propose a context module named CEM to address the low efficiency of rich channel information in backbone and exploit context features concurrently. More importantly, our module is lightweight and effective for RSOD, which is validated by experiments in Section IV.

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 $X \in \mathbb{R}^{C\times1\times1}$, for a MLP, the total number of parameters is $2 \times C \times C/r$, where r = 16 (refer to [28]); for a 1D convolution, the total number of parameters is k, where k is the kernel size of 1D convolution.
For example, the total params of ECA is 5×3 + 5×4 + 7×6 + 7×3 = 98.

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 AFPN for multi-scale object detection. The details of the three components and loss function are presented below.

A. Enhanced Effective Channel Attention

Channel attention mechanisms are efficient to help 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 (illustrated in Table II). To restrain the disturbance of complicated background, we design EECA mechanism inspired by ECA [39]. Its structure is shown in Fig. 2. Different from ECA, our EECA utilizes two 1D convolutional layers to capture non-linear 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.

Suppose 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: $X_{avg}, X_{max} \in \mathbb{R}^{C\times 1\times 1}$, which denote average-pooled features and max-pooled features, respectively. GAP and GMP can be expressed as

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

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

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 1D convolutional layers and a ReLU layer for local cross-channel correlation. EECA merges the output feature maps by using element-wise summation after exploiting each descriptor. Thus, the channel attention $M(X)$ is computed as

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

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

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

where $\otimes$ represents element-wise multiplication.

As for the setting of kernel size of 1D convolutional layers, we refer to the non-linear 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

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

where $|t|_{odd}$ denotes the nearest odd number of t. If t is an even number, then $|t|_{odd} = t+1$; otherwise, $|t|_{odd} = t$. 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.

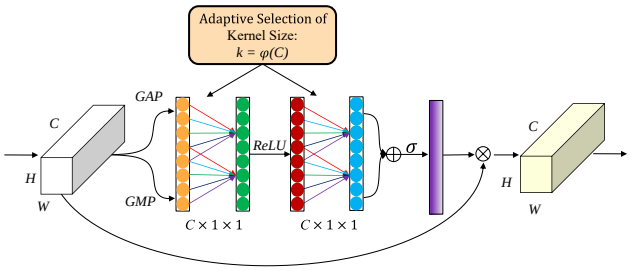


Fig. 2. Illustration of EECA. Given the aggregated features obtained by GAP and GMP, EECA generates channel attention by performing two fast 1D convolutions of size k with ReLU and Sigmoid functions.

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 amount of parameters and computational complexity. Experiments in Section IV reveal the superiority of EECA for RSIs.

B. Adaptive Feature Pyramid Network

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

The presented AFPN adopts the similar pipeline as Balanced FPN [21] and CE-FPN [20], which has a process of first aggregating multi-scale features and then splitting them into feature pyramids. Moreover, we put forward the Selective Refined Module (SRM) to perform adaptive balanced fusion between various spatial positions and channels, as shown in Fig. 3. To our best knowledge, AFPN can better extract the features of multi-scale 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 $C_4 \in \mathbb{R}^{C\times 2H\times 2W}$, $C_3 \in \mathbb{R}^{C\times 4H\times 4W}$, $C_2 \in \mathbb{R}^{C\times 8H\times 8W}$. Here, C represents the number of channels (equal to 256 in our algorithm), and H, W are 1/32 of the length and 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, C_5, respectively.

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

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

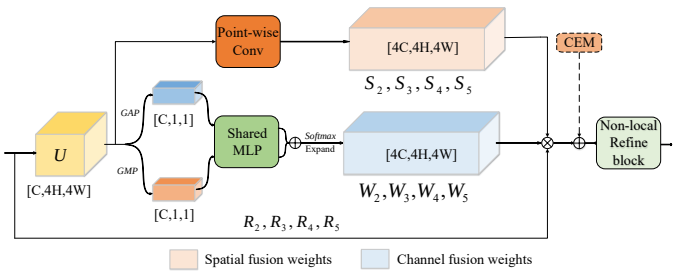


Fig. 3. Illustration of Selective Refined Module (SRM). It applies two parallel branches to capture spatial fusion weights and channel fusion weights of AFPN.

where $\oplus$ represents element-wise 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

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

We split W into four vectors $W_2, W_3, W_4, W_5 \in \mathbb{R}^{C\times1\times1}$ and expand them into $\mathbb{R}^{C\times4H\times4W}$, which represent 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 multi-scale object detection.

With respect to spatial fusion weights, SRM acquires it conveniently as illustrated in Fig. 3. SRM adopts a point-wise convolution to capture spatial fusion weights, which efficiently carries out spatial information interaction with low computation. The spatial fusion weights $S \in \mathbb{R}^{4C\times4H\times4W}$ is formed as

$$S = softmax(PointConv(U)), \quad (8)$$

where $PointConv(\cdot)$ indicates the processing of 1×1 point-wise convolutional layer with ReLU function. Besides, SRM applies $softmax$ function on the same spatial locations of different channels in S. We can also divide S into four vectors $S_2, S_3, S_4, S_5 \in \mathbb{R}^{C\times4H\times4W}$, which represent 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, W_5 and spatial fusion weights S_2, S_3, S_4, S_5, the refined integrated feature $\widetilde{U} \in \mathbb{R}^{C\times4H\times4W}$ can be generated as

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

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

AFPN utilizes 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 multi-scale features. The processing of generating the final detection maps of AFPN is described in the following subsection.

C. Context Enhancement Module

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 sub-pixel branches [53] to instill diverse spatial context information into integrated feature U. Theoretically, the spatial context information obtained by sub-pixel branches can reduce the loss in channels of F_5, thus improving final feature pyramid simultaneously. Sub-pixel 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 up-sampling as

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

where F^{SR}, 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 sub-pixel convolutional layers, the multi-scale spatial context information is acquired without much computational costs. Then CEM reconciles the channels of the spatial context information into C dimensions via convolutional layers with kernel of 1×1. Finally, the spatial context information is unified into a vector with the size of $C\times4H\times4W$ by ratio-invariant adaptive pooling. Considering the aliasing effect caused by adaptive pooling, we adopt ASFM rather than simple element-wise 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 multi-branch contextual features into the adaptive feature $\widetilde{U}$ eventually (refer to Eq. 9). Through such operations, CEM enriches the multi-scale 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 multi-scale 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 multi-task function is used to balance classification and localization task

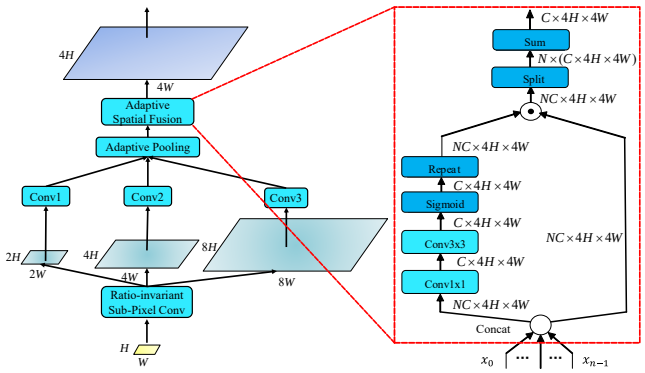


Fig. 4. The left one is the illustration of CEM, and the other is the illustration of ASFM.

loss [10]. In this paper, the total loss function is defined as

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

where L_{rpn}, L_{cls} and L_{loc} denote RPN loss, classification loss and 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

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

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

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

Here, the relationship between parameters γ, α, and b satisfies 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, etc. 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 10 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 size ranging from 533×597 to 1728×1028 pixels. According to [48], we split 75% of the dataset (879 images) as training set and 25% of it (293 images) as testing set.

3) DIOR [8]: It is the largest dataset for the horizontal object detection in geospatial RSIs. It includes 23463 images (800×800 pixels) with 192472 instances of 20 classes. This dataset is divided into 11725 images (50% of datatset) as training set and the remaining 11738 images as testing set.

B. Evaluation Metrics

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

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

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

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

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

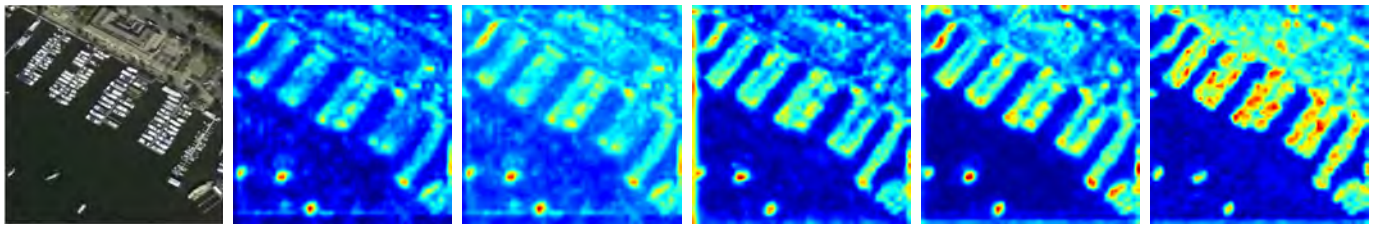
$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}, \tag{18}$$

respectively, where N_{TP}, N_{FP} and N_{FN} denote the number of true positives, false positives and false negatives, respectively. For a certain predicted b-box, the conditions for satisfying true positives are as follows. The first is that the IoU between the predicted b-box and a 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 PyTorch framework (PyTorch1.6) in Ubuntu 18.04 operating system, and all experiments are performed on 2 NVIDIA TITAN RTX GPUs with 24-GB memory for per GPU. ResNet50 [43] pretrained on ImageNet-1K classification task is the backbone. We adopt *kaiming-normal* to initialize new layers and stochastic gradient descent algorithm (SGD) to optimize the parameters of 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 8-*th* and 14-*th* epoch, the learning rate is reduced to 0.1 and 0.01 times, respectively. The size of input images for network is 800×800 for three datasets. The batch size of 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].

Baseline+eeca.png (b) baseline+se.png (b) heatmap_EECA.png (b) heatmap.png (b) baseline +eca.png



(a) input image (b) baseline (c) baseline + SE (d) baseline + CBAM (e) baseline + ECA (f) baseline + EECA

Fig. 5. Features visualizations of different channel attention mechanisms. (a) is the input image to network from NWPU VHR-10 dataset. (b)-(f) are the 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]. 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], BALANCED FPN [21] ON NWPU VHR-10 DATASET [7]. 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

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 1D convolution with $\gamma = 2$. In addition, "EECA w/o share" denotes that the GAP and GMP adopt two separate parallel paths instead of shared 1D 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 1D convolutional layers. "EECA $\gamma = 1$" has 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 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 1D convolutional layers with stronger non-linear 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 non-linear modeling between local cross-channel in CNNs is very crucial for remote sensing image 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 feature pyramid networks. 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 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 multi-scale feature maps of RSIs. The comparative FPNs [16], [17], [21] cannot address this problem well, which result in less significant performance improvements than our AFPN.

E. Ablation Study

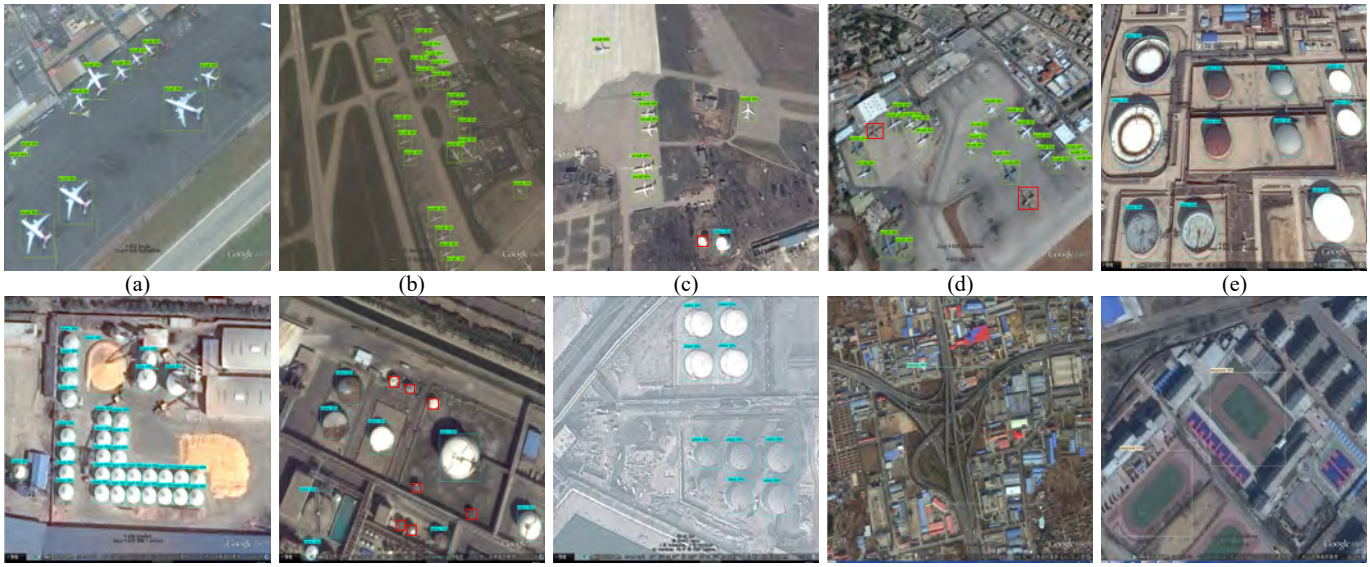
We set up ablation experiments on RSOD and NWPU VHR-10 datasets to prove the effects of EECA, AFPN and CEM in our algorithm. Table IV displays the results of ablation experiments.

1) Baseline setup: Our baseline is Faster RCNN [10] with FPN [15]. It uses ResNet50 as backbone, multi-scale RoI Align for regional feature extraction and Balanced Smooth L1 loss [21] as regression task loss. In all experiments, all hyperparameters are consistent for 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

TABLE IV
ABLATION STUDY ON NWPU VHR-10 DATASET [7] AND RSOD DATASET [40]. 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



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

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

TABLE V
COMPARISON WITH STATE-OF-THE-ARTS ON RSOD [40]. * DENOTES OUR IMPLEMENTATION AND 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**

background. In Table IV, our method (only with EECA) can reach 91.92% mAP (1.12%↑) and 91.90% mAP (0.99%↑) on NWPU VHR-10 dataset and 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 remote sensing images. 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.

3) Effect of AFPN: To detect multi-scale and dense objects efficiently in RSIs, AFPN is proposed. We compare it against baseline to validate the strategy of AFPN. Our method only with AFPN can obtain 92.37% mAP (1.57%↑) on NWPU VHR-10 dataset, and 92.29% mAP (1.38%↑) on RSOD dataset respectively. It illustrates that the simple strategy of spatial localities and channels of multi-scale feature maps via AFPN really contributes to the network for interpreting RSIs. Additionally, 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 lost 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 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 baseline, we observe that the detection performance is further improved, reaching 93.24% and 93.37% mAP on NWPU VHR-10 and RSOD, respectively. The reason is that CEM helps AFPN catch multi-scale semantic information from the deepest-level layer of backbone, which is remarkable for multi-scale object detection in RSIs.

TABLE VI
COMPARISON WITH STATE-OF-THE-ARTS ON NWPU VHR-10 [7]. * DENOTES OUR IMPLEMENTATION AND 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**

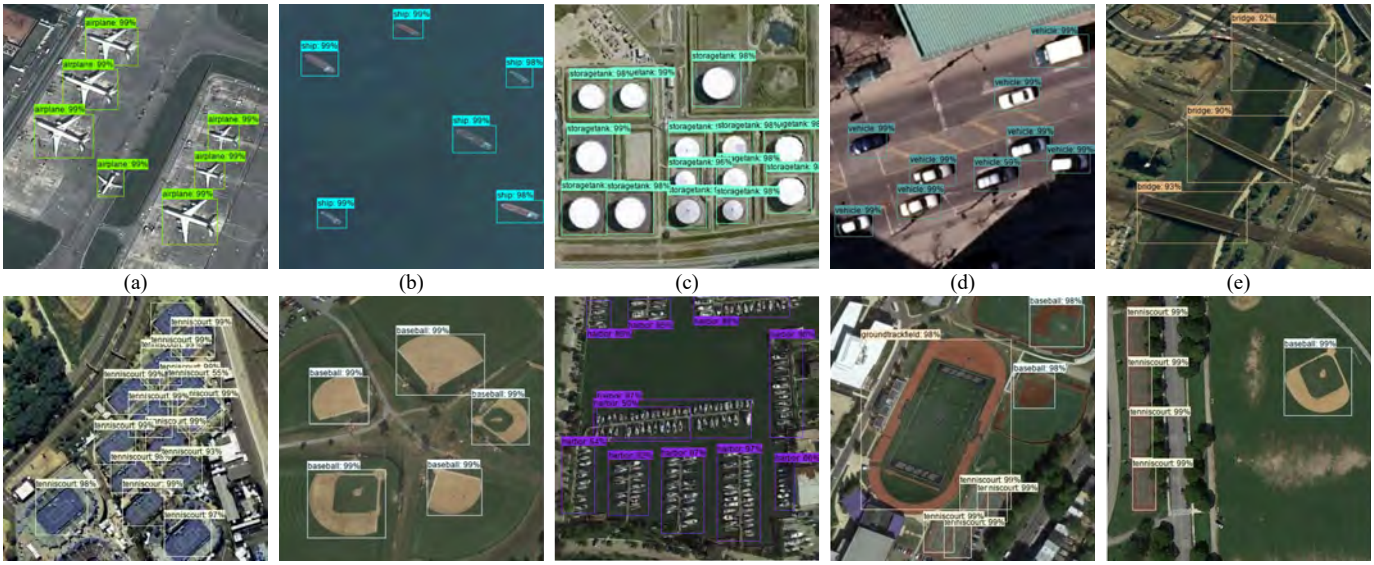


Fig. 7. Some representative detection results of our ABNet on NWPU VHR-10 dataset. (a) Airplane. (b) Ship. (c) Storage tank. (d) Vehicle. (e) Bridge. (f) Tennis court. (g) Baseball diamond. (h) Harbor. (i) Ground track field, Tennis court and Baseball diamond. (j) Tennis court and Baseball diamond.

TABLE VII
COMPUTATION TIME COMPARISON OF EIGHT APPROACHES ON 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

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

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

1) Results on RSOD: As displayed in Table V, ABNet reaches 94.17% mAP in 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 categories. By contrast with FPN (baseline), the ABNet performs better in all categories, and it upgrades the AP by a percentage 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 EECA mechanism. EECA suppresses negative information of complex background 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], RICAOD [48], Faster RCNN [10], FPN [15], SCRDet [35], CAD-Net [32]. In addition, we add the state-of-the-art anchor-free algorithm FCOS [57] to the comparison.

TABLE VIII
COMPARISON WITH STATE-OF-THE-ARTS ON DIOR [8]. * DENOTES OUR IMPLEMENTATION AND 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.

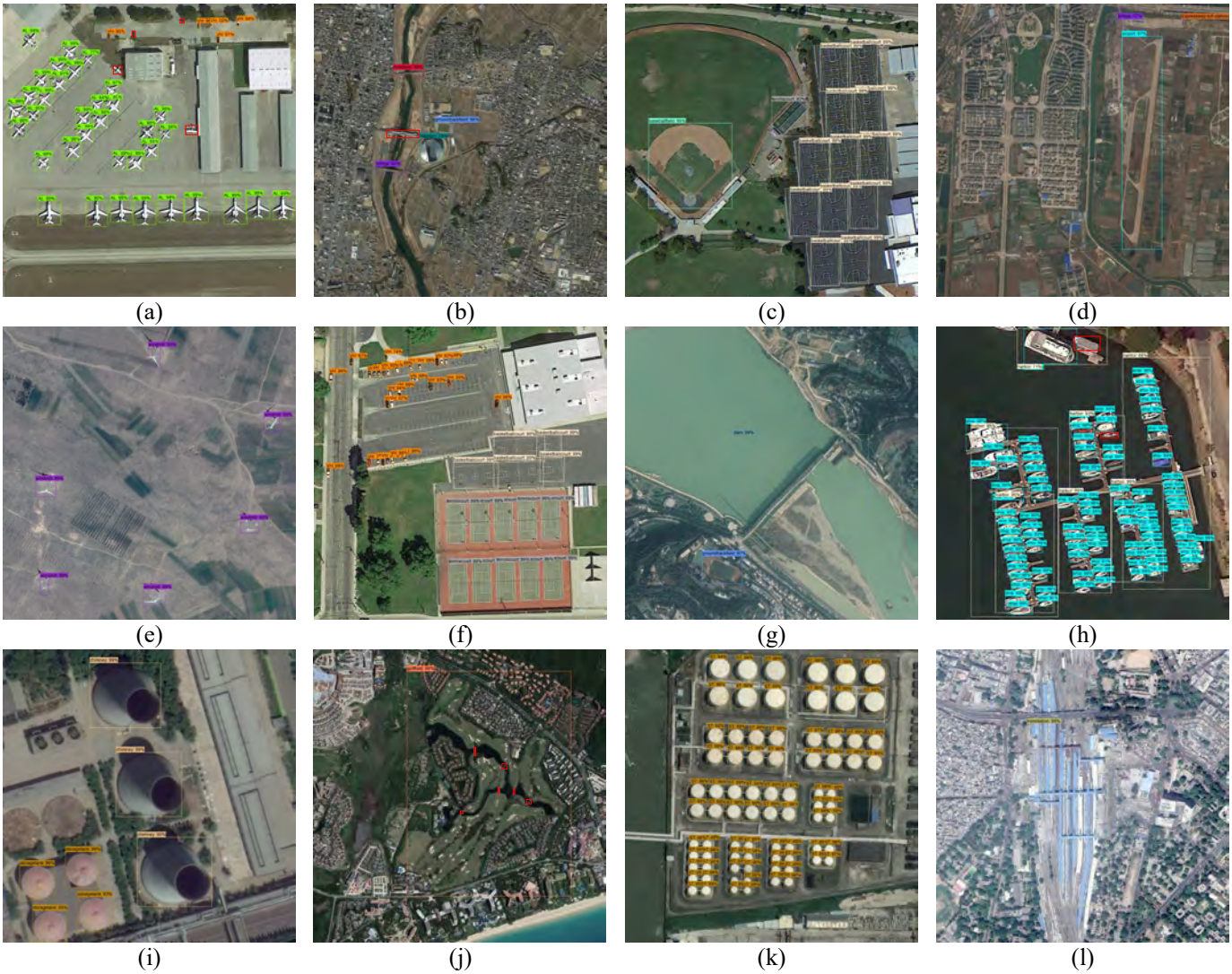


Fig. 8. Some representative detection results of ABNet on DIOR dataset. (a) Airplane and Vehicle. (b) Bridge, Stadium, Ground track field and Overpass. (c) Baseball filed, Tennis court and Basketball court. (d) Bridge, Airport and Expressway-toll-station. (e) Windmill. (f) Vehicle, Basketball court and 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, mAP_{50} OF DIFFERENT METHODS ON DIOR DATASET [8]. 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**

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, 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 18.18% increase in AP compared with FPN (baseline), which also shows the efficiency of 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 multi-scale object detection by constructing various submodules and combining with appropriate loss function.

Table VII shows the average computation time of eight approaches on NWPU VHR-10 dataset, which is widely adopted in remote sensing object detection [55]. As a fast one-stage detector, YOLOv3 [13] has the lowest running time among competitors. The computation time of ABNet has only a 0.02s gap with YOLOv3 and outperforms 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 trade-off because it uses ResNet50 as backbone instead of heavy ResNet101, and the proposed structures (EECA, AFPN, 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, 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 the 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, 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-size objects.

G. Qualitative Analysis

As shown in Figs. 6, 7 and 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 Figs. 6(a)-6(b) with various scales distribution can be well detected. Unfortunately, several small objects are missed in Figs. 6(c), 6(d) and 6(g). As shown in Fig. 8, ABNet can overcome the disturbance of complex background, multi-scale 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", "storage tank". However, ABNet fails to detect tiny objects as shown in Figs. 8(a) and 8(j).

[2]https://github.com/ultralytics/yolov3
[3]https://github.com/open-mmlab/mmdetection

V. Conclusion

In this paper, an improved detector ABNet with three upgrades based on Faster RCNN is proposed for RSIs. Firstly, 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. Secondly, AFPN is developed, which only introduces a MLP, a point-wise convolution and a Non-local block to integrate feature maps of various scales efficiently. Thirdly, CEM is deployed to combine the deepest-level features of 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 EECA mechanism in other remote sensing tasks.

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, 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, p. 107992, 2021.

[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, 2020.

[4] G. Ganci, A. Cappello, G. Bilotta, and C. Del, "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, p. 111426, 2020.

[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, 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, 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, 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, 2020.

[9] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021, early access, 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, 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)*, 2016, pp. 779–788.

[12] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6517–6525.

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

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

[15] 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)*, 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)*, 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)*, 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)*, 2020, pp. 12 595–12 604.

[19] M. Tan, R. Pang, and V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10 778–10 787.

[20] Y. Luo, X. Cao, J. Zhang, X. Cao, J. Guo, H. Shen, T. Wang, and Q. Feng, "CE-FPN: Enhancing channel information for object detection," *arXiv preprint arXiv:2103.10643*, 2021.

[21] J. Pang, K. Chen, Q. Li, Z. Xu, H. Feng, J. Shi, W. Ouyang, and D. Lin, "Towards balanced learning for instance recognition," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1376–1393, 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)*, 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*, 2021, early access, 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, 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, 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, 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, 2021.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 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.*, 2020, early access, 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)*, 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, 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. 10 015–10 024, 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.*, 2021, early access, 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, 2020.

[35] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 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)*, 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 im-

[49] X. Peng, I. Tsang, Y. Yang, and H. Tang, "Connection between nuclear-norm ..., *IEEE Trans. Geosci. Remote Sens.*, 2021, early access, doi: 10.1109/TGRS.2021.3068970.
... representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3218–3224, Aug. 2018.
[50] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional ... a sparse coding for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1823–1831.
[51] W. Luo, W. Li, and G. Fu, "An anomaly detection with sparse coding inspired ... channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11 531–11 539.
[52] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.
[53] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.
[54] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6823–6834.
[55] J. Y. Zhou *et al.*, "SC2Net: Sparse lstms for sparse coding," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4588–4595.
[56] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.
[57] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, "Anomaly detection in hyperspectral images based on low-rank and sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 1990–2000, Apr. 2016.
[58] W. Xie, B. Liu, Y. Li, J. Lei, C. Chang, and G. He, "Spectral adversarial feature learning for anomaly detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2352–2365, Apr. 2019.
[59] A. P. Bradley, "The use of the area under the ROC curve in the evaluation ... learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

Yunsong Li (Member, IEEE) received the M.S. ... degree in telecommunication and information systems, and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 1999 and 2002, respectively.

He joined the School of Telecommunications Engineering, Xidian University, in 1999, where he is a Professor. He is the Director of the Image Coding and Processing Center, State Key Laboratory of Integrated Services Networks, Xi'an. His research interests focus on image and video processing and high-performance computing.

[53] W. Shi, J. Caballero, F. Huszár, J. Totz, P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "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)*, 2016, pp. 1874–1883.

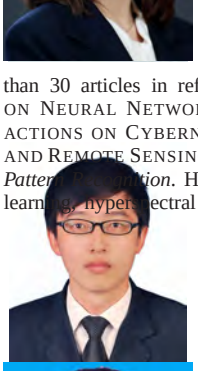
Tao Jiang received the B.E. degree in telecommunication engineering from Xidian University, Xi'an, China, in 2018, where she is pursuing the M.S. degree with the Image Coding and Processing Center, State Key Laboratory of Integrated Services Networks.

Her research is anomaly detection, hyperspectral ... and deep learning.

[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)*, 2019, pp. 9626–9635.

Wenfeng Lin (Member, IEEE) received the B.E. degree in computer ... degree in electronic information science and technology from the University of Jinan, Jinan, China, in 2011, the M.S. degree in communication and information systems from Sandhu University, Hangzhou, China, in 2014, and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2017.

He is currently an Associate Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. She has authored or coauthored more than 30 articles in refereed journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the AAAI Conference on Artificial Intelligence, and *Pattern Recognition*. Her research interests include neural networks, machine learning, hyperspectral image analysis, and high-performance computing.



Qiang Li received the B.E. degree in measurement ... & control technology and instrument from 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, China, in 2018.

He is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN). His research interests include hyperspectral image processing and computer vision.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE TRANSACTIONS ON IMAGE PROCESSING, and a Monograph in Springer. Her research interests include hyperspectral image analysis, scene understanding, and content-based image retrieval.

She was a Visiting Scholar with the Department of Computer Science, University of California at Los Angeles, Los Angeles, CA, USA, from 2014 to 2015. He is an Associate Professor with the School of Telecommunications Engineering, Xidian University, and a Member of the Image Coding and Processing Center, State Key Laboratory of Integrated Services Networks. His research interests focus on image and video processing, computer vision, and cus... computing for big-data applications.

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

She is currently a Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. She is an Adjunct Professor with the College of Surveying and Geoinformatics, Tongji University, Shanghai, China. 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. She was a recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS). She served as a 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 at Shanghai, China, in 2012. She served as an Associate Editor of 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 for the IEEE JSTARS. She is currently a member of the IEEE Periodicals Review and Advisory Committee.

Qi Wang (M'15-SM'15) 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 currently 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.