This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Lightweight Oriented Object Detection using Multi-scale Context and Enhanced Channel Attention in Remote Sensing Images

Qiong Ran, Qing Wang, Boya Zhao, Yuanfeng Wu, *Senior Member, IEEE*, Shengliang Pu, *Student Member, IEEE*, Zijin Li

***Abstract*—Object detection is a focal point in remote sensing applications. Remote sensing images typically contains a large number of small objects and a wide range of orientations across objects. This results in great challenges to small object detection approaches based on remote sensing images. Methods directly employ channel relations with equal weights to construct information features leads to inadequate feature representation in complex image small object detection tasks. Multi-scale detection methods improve the speed and accuracy of detection, while small objects themselves contain limited information, and the features are easily lost following down-sampling. During the detection, the feature images are independent across scales, resulting in discontinuity at the detection scale. In this paper, we propose the Multi-Scale Context and enhanced Channel Attention (MSCCA) model. MSCCA employs PeleeNet as the backbone network. In particular, the feature image channel attention is enhanced and the multi-scale context information is fused with multi-scale detection methods to improve the characterization ability of the convolutional neural network. The proposed MSCCA method is evaluated on two real datasets. Results show that for 512 × 512 input images, MSCCA was able to achieve 80.4% and 94.4% mAP on the DOTA and NWPU VHR-10, respectively. Meanwhile, the model size of MSCCA is 21% smaller than that of its predecessor. MSCCA can be considered as a practical lightweight oriented object detection model in remote sensing images.**

***Key words*—Object Detection, Multi-scale Context, Channel Attention, Lightweight convolutional neural network, Remote Sensing.**

I. Introduction

The object detection plays a key role in remote sensing algorithms and applications. They can be roughly divided into traditional and deep learning object detection approaches. Traditional object detection methods (HOG[1], SVM[2], DPM[3] etc.) generally include a region proposal, feature extraction and classification, resulting in a low detection efficiency and poor accuracy due to complex procedures, a large number of redundant windows and the poor robustness of manual feature extraction methods. Thus, traditional detection methods are hardly meeting the object detection performance demands. The emergence of deep learning-based methods has achieved significant breakthroughs in object detection [4]-[6]. Deep learning-based object detection methods mainly can be divided into two types: i) two-stage detection models, which defines detection as a "coarse-to-fine" process; and ii) one-stage detection models, which defines detection as a "one-step" process [7].

Two-stage detection approaches are generally region-based and extract a set of object proposals that potentially contain the objects using methods such as selective searching or region proposals. These sets are subsequently fed into a Convolutional Neural Network (CNN) for feature extraction. The classifiers then predict the presence of an object within each region and recognize the object categories. R-CNN[8] is a typical two-stage detector that generates proposals by selective searching and normalizes their size and inputs them to the CNN to extract the features. SVM is then applied to recognize object categories within each region. Fast R-CNN[9] improves R-CNN by using a multi-task loss to increase the detection quality. Faster R-CNN[10] introduces the Region Proposal Network (RPN), whereby the majority of the individual blocks in the object detection framework (region proposal, feature extraction, bounding box regression, etc.) are gradually integrated into an end-to-end learning framework. Mask R-CNN[11] includes a branch to segment an object based on Faster R-CNN and simultaneously performs instance segmentation and object detection. Libra R-CNN[12] integrates IoU-balanced sampling, a balanced feature pyramid and a balanced L1 loss to reduce the imbalance at the sampling, feature extraction and training procedures, respectively. Although two-stage object detection methods have made a great progress in detection tasks, they are limited by large amounts of parameters and slow detection speeds. HSP[13] considers the utilization and propagation of hierarchical semantic information in the optimized process of the detection network to improve object detection performance in remote sensing imagery.

One-stage detection methods apply a single CNN to divide the image into multiple regions and simultaneously predict the bounding boxes and category of each region. This process greatly improves the detection speed, yet reduces the detection accuracy compared to two-stage detectors. YOLO[14] is a typical one-stage object detection method that treats object detection as the solution of a regression problem, applying a single CNN to the full image. This network simultaneously predicts the bounding boxes and category for each region. SSD[15] is an additional one-stage detection method that sets default boxes with different aspect ratios in each feature map to

This work was supported by the National Natural Science Foundation of China under Grant 41871245, Grant 62001455. (*Corresponding author: Yuanfeng Wu*, wuyf@aircas.ac.cn)

Qiong Ran and Qing Wang are with the College of Information Science and Technology at Beijing University of Chemical Technology, Beijing, China.

Boya Zhao, Yuanfeng Wu, Shengliang Pu and Zijin Li are with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhaoby@aircas.ac.cn;wuyf@aircas.ac.cn;pusl@aircas.ac.cn)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

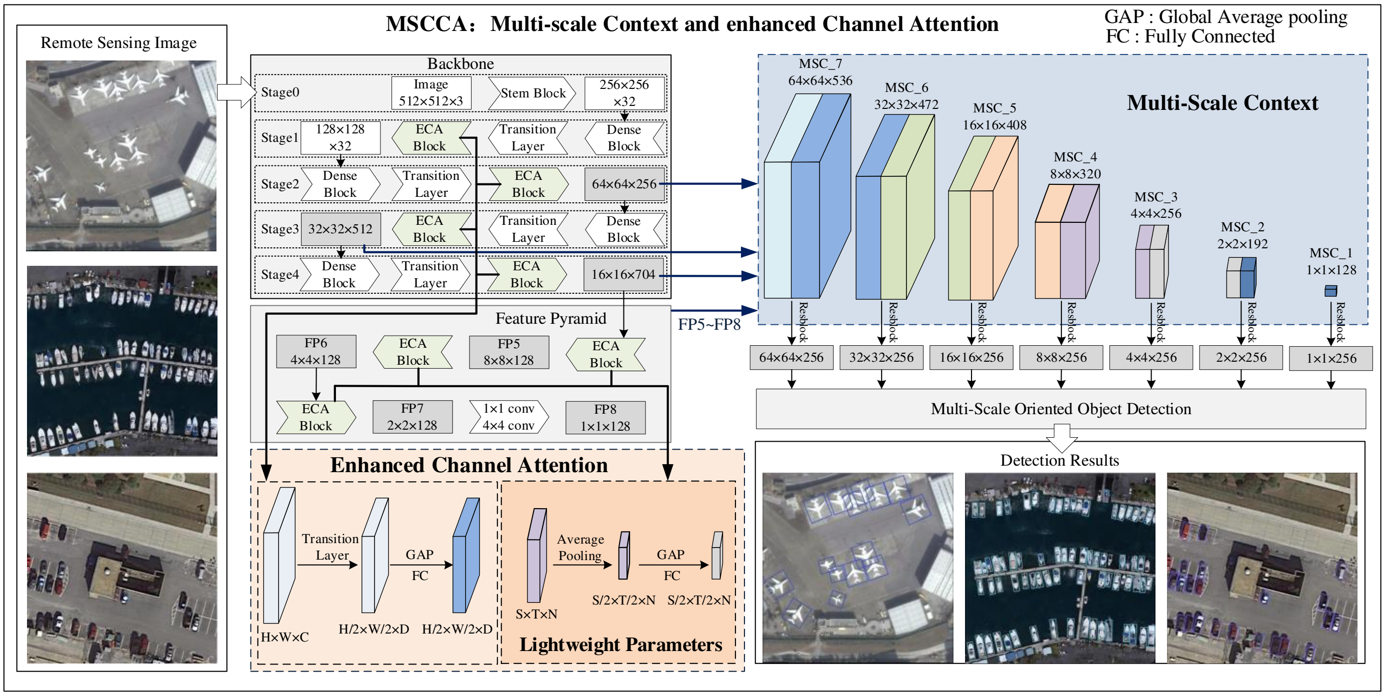

Fig. 1. The structure of MSCCA.

perform multi-scale detection, significantly improving the one-stage detector detection accuracy. FMSSD[16] leverages the Atrous Spatial Feature Pyramid (ASFP) module to integrate the context information into the framework, improving the robustness of features. RetinaNet[17] proposes the focal loss, whereby the detector pays more attention to samples that are difficult to classify during the training process. This maintains a high detection speed while matching the accuracy of two-stage detection methods. RefineDet[18] proposes the Anchor Refinement Module (ARM) and Object Detection Module (ODM) to improve the detection efficiency without reducing the detection speed. M2Det[19] propose the Multi-Level Feature Pyramid Network (MLFPN), and construct more object feature pyramids to detect objects at different scales. Based on FCN, FCOS[20] is an anchor-free detector that abandons anchor generation process, reducing memory footprints and improving the detection accuracy. MS-VANs [21] proposed a visual attention-based network and simultaneously predict object class at each pixel of the feature maps, and use visual attention network to highlight the features from the object region and decrease the influence of cluttered backgrounds. S^2A-Net [44] implemented full feature alignment and alleviates the inconsistency between regression and classification by using Feature Alignment Module (FAM) and Oriented Detection Module (ODM).

For small object detection tasks, SNIP[23] and SNIPER[24] employ scale normalization, and only detect objects with a fixed size for scale-specific feature maps. SNIPER reduces the computation of the multi-scale image pyramid generation and accelerates multi-scale training. DEFace[25] proposes the extended Feature Pyramid Network (FPN[26]) module with a Receptive Context Module (RCM) to enhance the distinguishability and robustness of features. TridentNet[27] constructs a parallel multi-branch architecture and adopts a scale-aware training scheme to specialize each branch by sampling the object instances of proper scales for training. SCRDet++[28] introduces denoising process to object detection, whereby instance-level denoising on the feature map is performed to enhance the detection of small and cluttered objects. Stitcher[29] dynamically generates stitched images to enrich small object samples and adaptively determines whether the input of the next iteration is the original or the stitched image, which improves the small object loss contribution.

In the traditional convolutional pooling process, the convolution operation does not consider the dependence of each feature channel. In addition, the importance of each channel in the generated feature image is considered to be the same, yet in the actual problem, the importance is actually distinct across channels. One-stage detection methods employ multi-scale detection that extract multi-scale feature maps from different layers of the network for predictions. Although this does not increase the amount of calculations, the small object itself has less pixel information and is easily lost during downsampling[30].

In this paper, we propose the Multi-Scale Context and enhanced Channel Attention (MSCCA) model. MSCCA employs PeleeNet as the backbone network. In particular, the feature image channel attention is enhanced and the multi-scale context information is fused with multi-scale detection methods to improve the characterization ability of the convolutional neural network[31]. The proposed method is evaluated on two real datasets. Results show that for 512 × 512 input images, MSCCA was able to achieve 80.4% and 94.4% mAP on the DOTA and NWPU VHR-10, respectively. Meanwhile, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

model size of MSCCA is 21% smaller than that of its predecessor. MSCCA can be considered as a practical lightweight oriented object detection model in remote sensing images.

The rest of this paper is structured as follows. In Section II, the multi-scale context and enhanced channel attention model is described. In Section III, two real datasets DOTA and NWPU VHR-10 are presented. In Section IV, the datasets are used to evaluate the proposed MSCCA model. Both the detection accuracy and model size are summarized. Section V concludes this paper with some remarks and hints at plausible future research lines.

II. Methods

Multi-scale context and enhanced channel attention model employs the PeleeNet [32] as the backbone, while the enhanced channel attention block is added to balance the channel features that have a positive effect on detection and weakens the channel features that have no effect. Then, the multi-scale context structure combines high-level and low-level features within the multi-scale detection framework. Fig. 1 presents the whole structure of MSCCA. Objects in remote sensing images typically exhibit large scale changes, arbitrary-orientation and irregular shapes. Thus, seven different scale feature maps are employed for multi-scale objects. Moreover, the quadrilateral representation is used in location loss for objects with arbitrary-orientation and irregular shapes.

A. Backbone

PeleeNet improves employs a large number of dense layers that consist of two branches that extract multi-scale features in the receptive field. ResBlock is added prior to the detection of each feature map. Moreover, MSCCA includes the ECA Block following each transition layer of the network structure. Due to the large size of the remote sensing images, in order to ensure the detection accuracy of small objects, the image is not resized and the input size set to 512 x 512 pixels.

The entire network consists of five stages. Stage0 only contains Stem Block, which is a low-cost and efficient module that can effectively improve the feature extraction ability with a minimal increase in computational cost. Stem Block initially employs a 3×3 convolution layer to downsample the image and subsequently divides it into two branches that use i) the max pooling layer to downsample the image and ii) one 1×1 and one 3×3 convolution layer. The two branches are merged to the channel dimension via concat.

The remaining components consists of dense and transition layers. The dense layer can acquire receptive fields at multiple scales and consists of two branches, one of which employs one 1×1 and one 3×3 convolution layer, while the other uses one 1×1 and two stacked 3×3 convolution layers. The two branches are merged with the previous feature to the channel dimension via concat. The transition layer includes a 1×1 convolution layer and a 2×2 average pooling layer with a stride of 2.

TABLE 1 BACKBONE

Stage		Layer	Feature Map
	Input		512×512×3
Stage0	Stem Block		256×256×32
Stage 1	Dense Block	Dense Layer x 3	128×128×32
	Transition Layer	1×1 Conv , stride 1	
		2×2 Ave pool ,stride 2	
	ECA Block		
Stage 2	Dense Block	Dense Layer x 3	64×64×256
	Transition Layer	1×1 Conv , stride 1	
		2×2 Ave pool ,stride 2	
	ECA Block		
Stage 3	Dense Block	Dense Layer x 3	32×32×512
	Transition Layer	1×1 Conv , stride 1	
		2×2 Ave pool ,stride 2	
	ECA Block		
Stage 4	Dense Block	Dense Layer x 3	16 ×16×704
	Transition Layer	1×1 Conv , stride 1	
		2×2 Ave pool ,stride 2	
	ECA Block		

B. Enhanced Channel Attention

The attention mechanism in the convolutional neural network draws on the human visual attention mechanism. Human vision quickly scans a global image to obtain the required object area, generally referred to as the focus of attention. Additional attention is then focused on this area to obtain more detailed information about the target object, while suppressing other useless information. In general, some features learned in the convolutional neural network will be redundant for the object detection task[33]. For example,the Relu layer will generate a large number of parameters with a value of 0, while visualizing the intermediate feature image can demonstrate the inability of some channels to detect the object. Thus, during network training, some channels are more important than other channels. In order to emphasize these important channels, we include the channel attention structure ECA Block to the model (Fig. 2) based on SE Block[34].

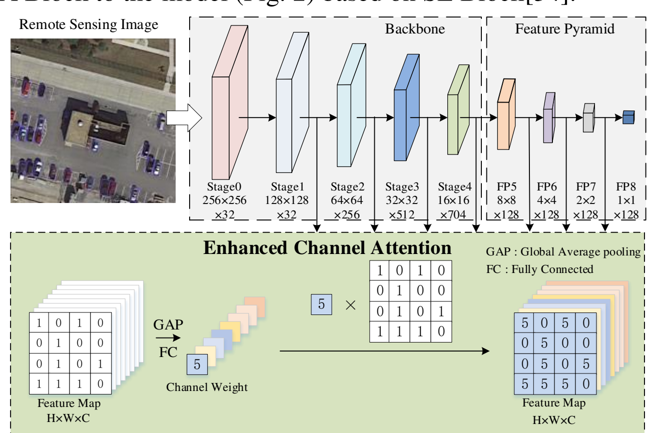

Fig. 2. Enhanced channel attention block.

In ECA Block, for any given feature map $X \in \mathbb{R}^{C\times H\times W}$, the global average pooling layer is implemented to generate features $M \in \mathbb{R}^{C\times 1\times 1}$:

$$M_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c\left(i,j\right) \tag{1}$$

where C is the number of channels; H and W are the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

height and width of the image, respectively; M_c indicates the feature obtained after the current channel performs global average pooling on X_c ; c is the current channel; and $X_c(i, j)$ is the feature value of input feature image X_c under coordinates (i, j) The feature image of each channel accumulates all the values, averages them to generate feature M_c and subsequently combines features M of C channels. Two Fully Connected (FC) layers are implemented. The first FC layer uses the Relu activation function to generate features of $C/r \times 1 \times 1$ size, where r is a hyperparameter and is used to change the ECA Block parameter in the network. Here, we set r to 16 following the previous experience of SE Block. The second FC layer uses the Sigmoid activation function to generate feature $S \in \mathbb{R}^{C \times 1 \times 1}$:

$$S = \sigma\left(W_2 \delta\left(W_1 M\right)\right) \tag{2}$$

where W_1 and W_2 represent two fully connected operations; δ and σ are two activation functions; S is the generated feature and represents the importance of each feature channel following feature selection. The normalized weight is multiplied to the feature of each channel to output feature $U \in \mathbb{R}^{C \times H \times W}$:

$$U_c = F_{scale}\left(X_c, S_c\right) = S_c X_c \tag{3}$$

where $U_c = [U_1, U_2, \ldots, U_c]$ represents the feature generated following the scale operation for current channel c . The scale operation multiplies each element in S_c and X_c to generate feature U for each channel.

We add the ECA Block to the proposed network to enhance the channel attention. The ECA Block is a simplified structure, which consists of a global average pooling operation, two full connections layers and a scale operation. Therefore, ECA Block can be used to replace the complex convolution component of the network in order to reduce the number of network parameters. For example, after replacing the additional convolution layer with ECA Block, the amount of network parameters is reduced from 7.06M to 5.08M. Our results demonstrate that including the ECA Block can generally improve the detection accuracy and reduce the amount of parameters (Section 4.3).

C. *Multi-scale Context*

The CNN in object detection is associated with a high shallow network resolution and low deep network resolution. Shallow convolution features represent the details of the object, while deep convolution features indicate the semantic information. However, using multi-scale feature maps for object detection ignores the detailed features in the shallow convolution features. Such shallow convolution features play a vital role in the detection of small objects. In order to fuse the scale context information[35][36], we include the FPN-based SC structure to the network. In table 2, we added convolutional layers to the end of the backbone to extract low-scale feature maps.

TABLE 2 FEATURE PYRAMID.

Feature Pyramid	Layer	Feature Map
FP5	1×1 conv , stride 1 3×3 conv , stride 2	8×8×256
FP6	1×1 conv , stride 1 3×3 conv , stride 2	4×4×256
FP7	1×1 conv , stride 1 3×3 conv , stride 2	2×2×256
FP8	1×1 conv , stride 1 4×4 conv , stride 1	1×1×256

MSCCA employs feature maps of different sizes to independently detect objects of varying sizes. In our proposed framework, the pyramid is constructed via bottom-up and top-down pathways, and lateral connections (Fig. 3). For every scale feature image (with the exception of the highest level), we upsample the spatial resolution by a factor of 2 (via bilinear interpolation upsampling) and merge with the same sized feature image convolved by 1×1. Feature maps of other sizes undergo the same procedure until a new feature pyramid is generated. Feature maps that fully integrate the scale context information are then adopted to detect objects of different scales.

$$U = \left[F_{upsample}\left(X\right) \oplus S\right] \tag{4}$$

For each feature layer X of the pyramid, $X \in \mathbb{R}^{C \times H \times W}$. X is upsampled to twice the scale and fused with S in the channel dimension to generate feature $U \in \mathbb{R}^{O \times H \times W}$. $S \in \mathbb{R}^{L \times H \times W}$ is a feature of the same scale as X . Fusing the features via the concat operation can make an excessively large feature dimension. Thus, we reduced the number of channels.

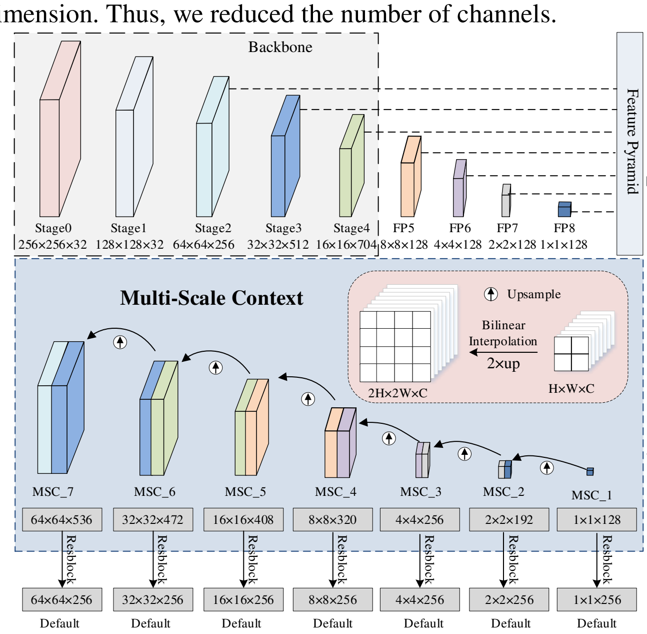

Fig. 3. Multi-scale context block

D. *Loss Function*

Remote sensing images typically exhibit arbitrary object orientations. Thus, MSCCA employs quadrilateral bounding boxes to detect objects across different directions. The location

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

information of the bounding box is expressed as (x, y, w, h) ,whereby (x, y) represents the center point coordinates of the bounding box, and w and h are the width and height of the bounding box, respectively. If we define the default box as $b=(x, y, w, h)$, then the corresponding quadrilateral is represented as $q=(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$, whereby (x_i, y_i) are the coordinates of the four vertices of the quadrilateral frame.

The loss function is divided into the confidence loss L_{conf} and the location loss L_{loc}:

$$L(x,c,l,g)=\frac{1}{N}\left(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)\right) \quad (5)$$

where α is the weight; N is the number of matched default boxes; $x\in\{1,0\}$ is the matching value indicating whether the default box matches the ground truth; c is the confidence; l is the predicted bounding box; and g is the ground truth. Positioning loss L_{loc} is a smooth L1 loss between the predicted bounding box and ground truth. If the overlap between the default box and ground truth exceeds the threshold (0.5), then it is considered as a positive sample:

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m}x_{ij}^{k}smooth_{L1}\left(l_i^m-\hat{g}_j^m\right) \quad (6)$$

where $m\in\{x, y, w, h, x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4\}$, (x, y) represents the center coordinates of the box; $x_{ij}^{k}\in\{1,0\}$ is an indicator of the match between the i-th predicted bounding box and the j-th ground truth; Pos is a positive sample; k is a ground truth object category; and $\hat{g}_j^m$ represents the coded ground truth, which ensures that the weight of the ground truth center position and weakens the width and height widths:

$$\hat{g}_j^x=\left(g_j^x-d_i^x\right)/d_i^w \qquad \hat{g}_j^y=\left(g_j^y-d_i^y\right)/d_i^h \quad (7)$$

$$\hat{g}_j^w=\log\left(\frac{g_j^w}{d_i^w}\right) \qquad \hat{g}_j^h=log\left(\frac{g_j^h}{d_i^h}\right) \quad (8)$$

$$\hat{g}_j^{x_1}=\left(g_j^{x_1}-d_i^{x_{min}}\right)/d_i^w \;\; \hat{g}_j^{y_1}=\left(g_j^{y_1}-d_i^{y_{min}}\right)/d_i^h \quad (9)$$

$$\hat{g}_j^{x_2}=\left(g_j^{x_2}-d_i^{x_{max}}\right)/d_i^w \;\; \hat{g}_j^{y_2}=\left(g_j^{y_2}-d_i^{y_{min}}\right)/d_i^h \quad (10)$$

$$\hat{g}_j^{x_3}=\left(g_j^{x_3}-d_i^{x_{max}}\right)/d_i^w \;\; \hat{g}_j^{y_3}=\left(g_j^{y_3}-d_i^{y_{max}}\right)/d_i^h \quad (11)$$

$$\hat{g}_j^{x_4}=\left(g_j^{x_4}-d_i^{x_{min}}\right)/d_i^w \;\; \hat{g}_j^{y_4}=\left(g_j^{y_4}-d_i^{y_{max}}\right)/d_i^h \quad (12)$$

where d represents the default box; d_i^w and d_i^h are the width and height of the default box, respectively; (x_{min}, y_{min}) and (x_{max}, y_{max}) represent the coordinates of the upper left and lower right points of the horizontal default box, respectively; and is the $smooth_{L1}$ loss defined as:

$$Smooth_{L1}(x)=\begin{cases}0.5x^2 & if\ |x|<1\\ |x|-0.5 & otherwise\end{cases} \quad (13)$$

Confidence loss L_{conf} is described in formula (17) and can be divided into the cross entropy loss of the positive and negative samples.

$$L_{conf}(x,c)=-\sum_{i\in pos}^{N}x_{ij}^{p}log\left(\hat{c}_i^p\right)-\sum_{i\in neg}log\left(\hat{c}_i^0\right)$$
$$\text{where } \hat{c}_i^p=\frac{exp\left(c_i^p\right)}{\sum_p exp\left(c_i^p\right)} \quad (14)$$

where c_i^p is the multi-category output; confidence $\hat{c}_i^p$ is obtained following the activation of the Softmax function; p represents the p-th category; the 0th category is the background; i is the i-th predicted bounding box; and Pos and Neg indicate positive and negative samples, respectively. In order to ensure a balance, the ratio of the positive to negative sample is set to 3:1.

III. Datasets

A. DOTA

The DOTA dataset [37] was published on CVPR by Wuhan University. DOTA is a large-scale dataset used for the object detection of aerial images. It contains 2,806 aerial images from different sensors and platforms. The images in the DOTA-v1.0 dataset were collected from Google Earth, some of which were taken by the satellite JL-1, and others were taken by the satellite GF-2 of the China Resources Satellite Data and Application Center. The size of each image ranges from approximately 800 × 800 to 4000 × 4000 pixels, and contains objects of various proportions, orientations, and shapes. Current object detection methods generally divide small objects into two categories: i) objects smaller than 32 × 32 pixels; and ii) objects with a width and height less than one-tenth of the original image. Fig. 4 presents the area distribution of all object types in the DOTA dataset, where the horizontal axis represents the object pixel area size and the vertical axis is the percentage of each category in a certain scale range. The DOTA dataset contains a large number of small objects, the majority of which are aircrafts, cars, and boats. The objects are divided into the following 15 categories: plane, ship, storage tank, baseball field, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field and swimming pool.

Due to the large number of pictures in the dataset and the large scale changes, we crop the pictures to a size of 512 × 512 pixels and randomly select 3/5 of the samples as the training set, 1/5 as the verification set, and 1/ 5 as the test set.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE 3 DOTA DATASET DETECTION RESULTS.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP	Fps
SSD[15]	41.0	24.3	4.55	17.1	15.9	7.72	13.2	39.9	12.0	46.8	9.09	30.8	1.36	3.50	0	17.8	59
YOLOv2[39]	76.9	33.8	22.7	34.8	38.7	32.0	52.3	61.6	48.5	33.9	29.2	36.8	36.4	38.2	11.6	39.2	30
RetinaNET[17]	78.2	53.4	26.3	42.2	63.6	52.6	73.1	87.1	44.6	57.9	18.0	51.0	43.3	56.5	7.4	50.3	14
R-FCN[40]	81.0	58.9	31.6	58.9	49.7	45.0	49.2	68.9	52.0	67.4	41.8	51.4	45.1	53.3	33.8	52.5	9
YOLOv3 错误!未找到引用源。	79.0	77.1	33.9	68.1	52.8	52.2	49.8	89.9	74.8	59.2	55.5	49.0	61.5	55.9	41.7	60.0	13
DSSD[42]	**91.1**	71.8	54.6	66.4	79.0	77.2	87.5	87.6	52.1	69.7	38.0	72.6	75.4	59.4	28.9	67.4	9
DYOLO[43]	86.0	71.4	54.6	52.5	79.2	80.6	87.8	82.2	54.1	75.0	51.0	69.2	66.4	59.2	51.3	68.1	17
FPN[26]	88.7	75.1	52.6	59.2	69.4	78.8	84.5	90.6	81.3	82.6	52.5	62.1	76.7	66.3	60.1	72.0	6
FMSSD[16]	89.1	81.5	48.2	67.9	69.2	73.5	76.8	90.7	82.6	73.3	52.6	67.5	72.3	80.5	60.1	72.4	16
DRN[44]	89.7	82.3	47.2	64.1	76.2	74.4	85.8	90.5	86.1	84.8	57.6	61.9	69.3	69.6	58.4	73.2	9.8
R^3Det [45]	89.4	81.1	50.5	66.1	70.9	78.6	78.2	**90.8**	85.2	84.2	61.8	63.7	68.1	69.8	67.1	73.7	10
SCRDET++[28]	90.0	84.3	55.4	73.9	77.5	71.1	86.0	90.6	87.3	87.0	**69.6**	68.9	73.7	71.2	65.0	76.8	13
FR-EST[46]	89.7	**85.2**	55.4	77.7	**80.2**	83.7	87.5	90.8	**87.6**	86.9	65.6	68.7	71.6	79.9	66.2	78.4	—
S^2A-NET[22]	89.2	84.1	56.9	79.2	80.1	82.9	**89.2**	90.8	84.6	**87.6**	71.6	68.2	**78.5**	78.2	65.5	79.1	34
Pelee	87.6	72.9	52.8	73.7	73.5	77.9	76.3	90.0	80.7	74.4	40.7	68.0	71.7	79.6	83.7	74.0	29.2
MSCCA	89.7	84.9	**64.5**	**81.3**	77.3	**83.9**	84.8	90.4	86.2	77.1	54.8	**79.7**	78.0	**84.4**	**89.1**	**80.4**	**32.3**

TABLE 4 NWPU VHR-10 DATASET DETECTION RESULTS.

Method	PL	SH	ST	BD	TC	BC	GT	HA	BR	VH	mAP
RICNN[47]	88.3	77.3	85.2	88.1	40.8	58.4	86.7	68.6	61.5	71.1	72.6
COPD[48]	89.1	81.7	97.3	89.3	73.2	73.4	82.9	73.3	62.8	83.3	80.6
Faster R-CNN[10]	94.6	82.3	65.3	95.5	81.9	89.7	92.4	72.4	57.5	77.8	80.9
HyperNet[49]	99.4	89.7	**98.6**	90.9	90.6	90.3	89.2	80.3	68.9	88.6	88.7
Pelee	99.5	**93.4**	90.8	**97.2**	90.7	96.0	95.9	88.9	**88.9**	90.7	93.2
MSCCA	**99.7**	90.4	90.8	90.8	**90.8**	**98.6**	**98.3**	**90.3**	88.2	**98.3**	**94.4**

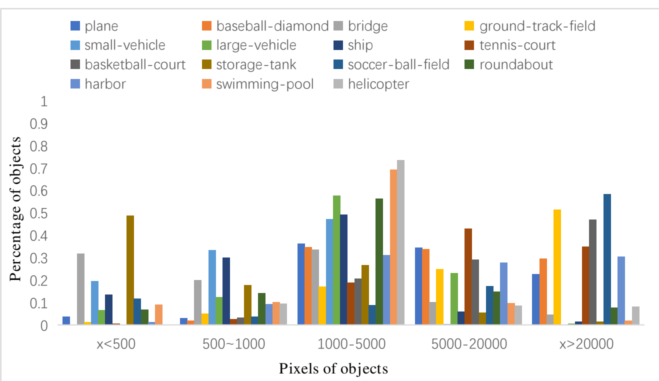

Fig. 4. Area distribution map of the objects contained in the DOTA dataset

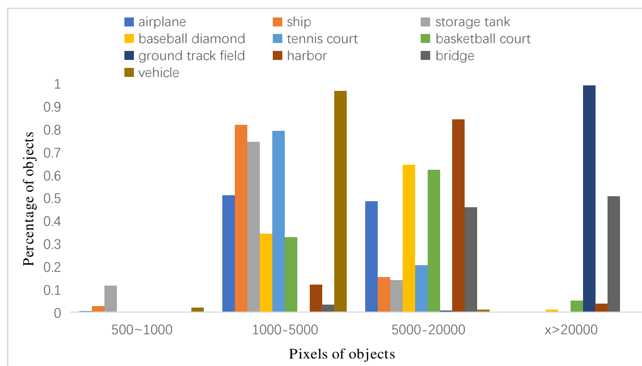

Fig. 5. Area distribution map of the objects contained in the NWPU VHR-10 dataset

B. NWPU VHR-10

The NWPU VHR-10 dataset [38] is derived from a 10-level geographic remote sensing dataset for space object detection. The dataset includes 650 images containing objects and 150 background images. The image content and object types/characteristics are similar to those of the DOTA dataset (Fig. 5). Although the dataset contains many object types, the number of samples is small, and the number and proportion of small objects is much less than that of the DOTA dataset. In particular, the NWPU VHR-10 dataset has almost no objects with an area of less than 1,000 pixels. The 10 types of objects are: airplane, ship, storage tank, baseball field, tennis court, basketball court, ground track field, harbor, bridge and vehicle.

IV. RESULTS

A. DOTA Dataset Results

The multi-scale object detection method generate candidate regions of different scales on the feature maps, which are of different receptive field sizes and the size of the default box is based on these receptive field sizes. The default box setting contains two features: the scale and aspect ratio. The scale of each feature image default box is set as follows:

$$S_k = S_{min} + \frac{S_{max} - S_{min}}{m-1}(k-1), k \in [1, m] \quad (15)$$

where S_k is the scale of the default box to the image; S_{min} and S_{max} represent the ratio of the lowest and highest scales, set to 0.15 and 0.9, respectively; and m is the number of feature maps of different sizes. Once the default box scale S_k of each feature image layer is determined, the specific default box is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

calculated according to the pre-defined aspect ratio. When the aspect ratio is 1, the side lengths of the two square default boxes are equal to S_k and $S_k^{'} = \sqrt{S_k S_{k+1}}$, where S_{k+1} is the default box scale of the feature image in the next layer. If the aspect ratio does not equal 1, the default box is calculated as follows

$$w_k^a = S_k \sqrt{a} \qquad h_k^a = S_k / \sqrt{a} \tag{16}$$

w_k^a and h_k^a are the width and height of the candidate region of the k-th feature image; and a is the value of the aspect ratio.

For the size of input image is 512×512, we select seven feature image scales to cover the different object sizes, as same as SSD[15], DSSD[42], FSSD[52], Rainbow SSD[53], Pelee[32], etc. The default box aspect ratios set to [[1,2,1/2], [1,2 ,3,1/2,1/3], [1,2,3,1/2,1/3], [1,2,3,1/2,1/3], [1,2,3 ,1/2,1/3], [1,2,1/2], and [1,2,1/2]].

During training, the pre-trained model is employed to initialize the parameters. The learning rate is set at 0.005 for the first 120,000 iterations and is subsequently reduce by an order of magnitude after every 40000 iterations computation, with 200,000 iterations in total. The momentum, weight decay and batch size are set to 0.9, 0.0005, and 16, respectively. The model is trained using the stochastic gradient descent method on four Nvidia Titan Xp GPUs.

Table 3 is the test results of the MSCCA model on the DOTA dataset, while Fig. 6 depicts the results of the model leaflet test. The result proves that MSCCA has higher detection accuracy than Pelee in detecting various objects.Pelee achieves a detection accuracy of 74% on the DOTA dataset, while that of the proposed MSCCA is 80.4%. This demonstrates the ability of the ECA Block and scale context features to improve the detection accuracy. The proposed MSCCA has a higher detection accuracy than S2A-NET.

The following is the detection results of SSD, Pelee and MSCCA on DOTA. As shown in Fig. 6, the information in Pelee is not enough to detect the objects. The prediction result of MSCCA outperforms the Pelee by a large margin. And the boxes of objects are regressed more accurately.

B. NWPU VHR-10 Dataset Results

The cropping, sample selection and settings of the NWPU VHR-10 dataset[38] follow those of the DOTA dataset. Seven feature maps of different sizes are used, and six default boxes of varying ratios are generated for each pixel and scale feature layer. However, in contrast to the DOTA dataset, the NWPU VHR-10 dataset only contains a horizontal manual annotation box, and thus the results are maintained in the horizontal box.

The pre-trained PeleeNet model is employed to initialize the parameters during training, with a 0.005 learning rate for the first 60,000 iterations that is subsequently reduced by an order of magnitude until the total 80,000 iterations are complete. The momentum, weight decay and batch size are set to 0.9, 0.0005, and 16, respectively and training is performed using the stochastic gradient descent method using four Nvidia Titan Xp GPUs.

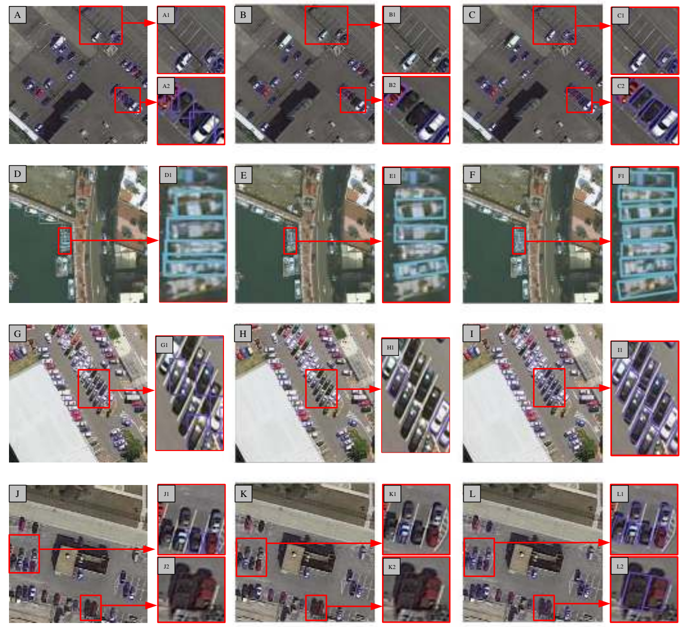

Fig. 6 . DOTA results of SSD (left), Pelee (middle) and MSCCA (right)

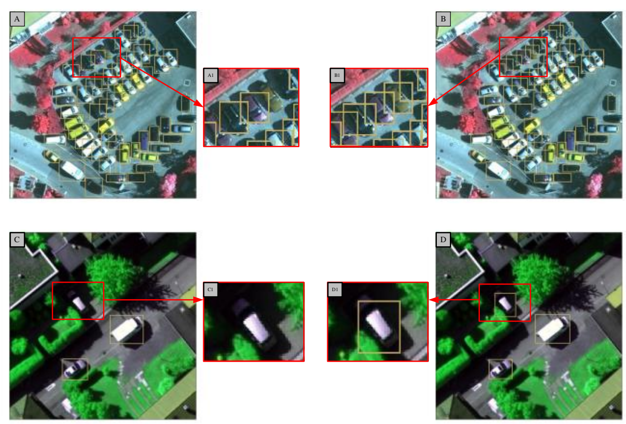

Fig. 7. NWPU VHR-10 results of Pelee (left) and MSCCA (right)

Table 4 is the detection results of the MSCCA model and other methods on the NWPU VHR-10 dataset. MSCCA outperforms the HyperNet by 5.7%. The following is the detection results of Pelee and MSCCA on NWPU VHR-10. Since there is no manual quadrilateral annotation in NWPU VHR-10 dataset, we use horizontal default box to detect.As same as results on DOTA, the detection effect of MSCCA is better than that of Pelee.

TABLE 5 INFLUENCE OF THE ECA BLOCK AND MSC STRUCTURE ON THE DETECTION PERFORMANCE FOR DOTA DATASET.

Resblock-> ECA	Transition layers-> ECA	Additional layers-> ECA	Add ECA after each stage	MSC	mAP	Model size	Speed (Nvidia Titan Xp)	Parameters	FLOPs
					74.0	26.5 MB	29.2 fps	6.56M	4.58G
√			√	√	75.4	22.1 MB	32.2 fps	5.48M	3.93G
			√		75.8	26.9 MB	30.0 fps	6.67M	4.58G
	√			√	78.9	25.2 MB	33.2 fps	6.26M	4.43G
				√	79.2	28.2 MB	31.9 fps	7.00M	5.36G
			√	√	80.2	28.6 MB	31.9 fps	7.10M	5.36G
		√	√	√	80.4	20.6 MB	32.3 fps	5.12M	5.02G

TABLE 6 INFLUENCE OF THE ECA BLOCK AND MSC STRUCTURE ON THE DETECTION PERFORMANCE FOR NWPU-VHR DATASET.

Resblock-> ECA	Transition layers-> ECA	Additional layers-> ECA	Add ECA after each stage	MSC	mAP	Model size	Speed (Nvidia Titan Xp)	Parameters	FLOPs
					93.2	26.4 MB	29.1 fps	6.52M	4.55G
√			√	√	93.5	21.9 MB	32.0 fps	5.43M	3.91G
			√		93.8	26.8 MB	30.0 fps	6.63M	4.55G
	√			√	93.8	25.0 MB	32.8 fps	6.22M	4.40G
				√	94.0	28.2 MB	32.0 fps	6.95M	5.33G
			√	√	94.3	28.4 MB	32.0 fps	7.06M	5.33G
		√	√	√	94.4	20.5 MB	32.4 fps	5.08M	4.99G

C. *Ablation Study*

In order to investigate the impact of the ECA Block and MSC structure on the detection results, we created several training models for the DOTA dataset and NWPU VHR-10 dataset to test using Nvidia Titan Xp and applied on Jetson TX2. We then evaluated the model size, detection speed and computational complexity of the proposed method.

Table 5 and Table 6 is the impact of each structure in terms of the detection accuracy, parameter file size, and detection speed under a single Nvidia Titan Xp GPU. In Table 5, without any structure, Pelee achieves a detection result of 74.0% mAP. Following the addition of the ECA Block after each network stage, the accuracy improves to 75.8% mAP. This demonstrates the ability of the ECA Block to strengthen the characterization performance of the network, thus improving the detection results. The inclusion of the MSC structure fusion scale context further improves the detection accuracy to 80.2% mAP. We then evaluate the impact of the ECA Block, replacing the complex convolutional layer in the network. Replacing the ECA Block with Resblock or the transition layer reduces the network parameters yet the detection accuracy is also weakened. Following this, we include a convolutional layer to provide small-scale feature maps for the multi-scale detection framework. We use a pooling layer to replace the convolutional layer downsampling, and subsequently add the ECA Block to enhance the channel attention.

For lightweight network, flops, model parameters and Memory Access Cost (MAC)[50][51] is widely used to measure the computational cost. Follow design guide of lightweight network, in the proposed structure MSC and ECA Block, we balanced the number of input and output channels for 1×1 convolution and make their ratio approach 1:1. This operation have been proved to reduce the MAC of the network. With the addition of MSC structure, the parameters and flops of the model are increased, but the inference time of the model is accelerated. Compared with Pelee, this structure was able to achieve a mAP of 80.4% and 6.4% higher than Pelee. The model size reduced from 26.5 MB to 20.6 MB and detection speed increased from 30.0 fps to 32.3 fps. Thus, the MSCCA can be considered as a lightweight oriented object detection model in remote sensing images.

V. Conclusions

A lightweight Multi-Scale Context and enhanced Channel Attention (MSCCA) model was proposed in this paper. It employs PeleeNet as the backbone network. The feature image channel attention is enhanced and the multi-scale context information is fused with multi-scale detection methods to improve the characterization ability of the convolutional neural network. Results show that for 512 × 512 input images, MSCCA was able to achieve 80.4% and 94.4% mAP on the DOTA and NWPU VHR-10, respectively. Meanwhile, the model size of MSCCA is 21% smaller than that of its predecessor. MSCCA can be considered as a practical lightweight oriented object detection model in remote sensing images. In the future, the proposed MSCCA model will be applied to edge devices for object detection application in remote sensing images. Moreover, computing optimization methods (like TenserRT) will be used to improve the processing efficiency of model inference procedures.

References

[1] N.Dalal and B.Triggs, "Histograms of oriented gradients for human detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., San Diego, CA, USA, pp. 886-893, July 2005.

[2] T. Malisiewicz, A. Gupta and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," Proc. IEEE Int. Conf. Comput. Vis., Barcelona, Spain, pp. 89–96, Nov 2011.

[3] P . F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep 2009.

[4] Mahdianpari M , Ghanbari H , Mohammadimanesh F , et al. A Meta-Analysis of Convolutional Neural Networks for Remote Sensing Applications[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, PP(99):1-1.

[5] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, Bing Zhang. “More Diverse Means Better: Multimodal Deep Learning Meets Remote Sensing Imagery Classification, IEEE Transactions on Geoscience and Remote Sensing,” 2020.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

[6] Danfeng Hong, Naoto Yokoya, Gui-Song Xia, Jocelyn Chanussot, Xiao Xiang Zhu. " X-ModalNet: A Semi-Supervised Deep Cross-Modal Network for Classification of Remote Sensing Data," ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 167: 12-23.

[7] Zou Z, Shi Z, Guo Y, et al, "Object detection in 20 years: A survey," 2018, [online] http://arxiv.org/abs/1905.05055v1.

[8] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and se-mantic segmentation," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, pp. 580–587, June 2014.

[9] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. Comput. Vis. Santiago, Chile, pp. 1440–1448, Dec 2015.

[10] S. Ren, K. He, R. Girshick and J. Sun, "Faster r-cnn:Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 39, no. 6, pp. 91–99, June 2016.

[11] K. He, G. Gkioxari, P . Dollár and R. Girshick, "Mask r-cnn," Proc. IEEE Int. Conf. Comput. Vis., Venice, Italy, pp. 2980–2988, Dec 2017.

[12] Pang J , Chen K , Shi J , et al, "Libra R-CNN: Towards Balanced Learning for Object Detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, pp. 821-830, June 2019.

[13] C. Xu, C. Li, Z. Cui, T. Zhang and J. Yang, "Hierarchical Semantic Propagation for Object Detection in Remote Sensing Imagery," Proc. IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 6, pp. 4353-4364, June 2020.

[14] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, pp. 779–788, June 2016.

[15] W. Liu, D. Anguelov , D. Erhan, C. Szegedy , S. Reed, C.-Y .Fu and A. C. Berg, "SSD: Single shot multibox detector," Proc. Eur. Conf. Comput. Vis., Amsterdam, The Netherlands, pp. 21–37, June 2016.

[16] Wang P , Sun X , Diao W , et al, "FMSSD: Feature-Merged Single-Shot Detection for Multiscale Objects in Large-Scale Remote Sensing Imagery," IEEE Trans. Geosci. Remote Sens., vol. 58, no. 9, pp.3377-3390, Dec 2019.

[17] Lin T Y , Goyal P , Girshick R , et al, "Focal Loss for Dense Object Detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp.2999-3007, July 2018.

[18] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li, "Single-shot refinement neural network for object detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USApp. 4203-4212, June 2018.

[19] Zhao Q , Sheng T , Wang Y , et al, "M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network," Proc. IEEE AAAI., New Orleans, LA, USA, pp. 9259-9266, February 2018.

[20] Tian Z, Shen C, Chen H, et al, "Fcos: Fully convolutional one-stage object detection," Proc. IEEE Int. Conf. Comput. Vis., Venice, Italy, pp. 9627-9636, Oct 2017.

[21] C. Wang, X. Bai, S. Wang, J. Zhou and P. Ren, "Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images," Proc. IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 2, pp. 310-314, Feb 2019.

[22] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. “Dynamic anchor learning for arbitrary oriented object detection,” arXiv preprint arXiv:2012.04150, 2020.

[23] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–snip," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, pp. 3578–3587, June 2018.

[24] B. Singh, M. Najibi and L. S. Davis, "Sniper: Efficient multi-scale training," Proc. Adv. Neural Inf. Process. Syst., Montréal, Canada, pp. 9333–9343, 2018.

[25] T. M. Hoang, G. P. Nam, J. Cho and I. -J. Kim, "DEFace: Deep Efficient Face Network for Small Scale Variations," IEEE Access, vol. 8, pp. 142423-142433, July 2020.

[26] Lin T Y , Dollar P , Girshick R , et al, "Feature Pyramid Networks for Object Detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, pp. 936-944, July 2017.

[27] Y . Li, Y . Chen, N. Wang and Z. Zhang, "Scale-aware trident networks for object detection," Proc. IEEE Int. Conf. Comput. Vis., Seoul, South Korea, pp. 6053-6062, October 2019.

[28] Yang X , Yan J , Yang X , et al, "SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing," 2020, [online] https://arxiv.org/abs/2004.13316.

[29] Chen Y , Zhang P , Li Z , et al, "Stitcher: Feedback-driven data provider for object detection," CVPR, 2020.

[30] J. Su, J. Liao, D. Gu, Z. Wang and G. Cai, "Object Detection in Aerial Images Using a Multiscale Keypoint Detection Network," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 1389-1398, Dec. 2021.

[31] Y. Li, Q. Huang, X. Pei, Y. Chen, L. Jiao and R. Shang, "Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 2148-2161, Dec. 2021.

[32] Wang, R.J., Li, X., Ao, S., Ling and C.X, "Pelee: A real-time object detection system on mobile devices," Proc. Adv. Neural Inf. Process. Syst., Montréal, Canada, pp. 1967-1976, Dec 2018.

[33] J. Wang, Y. Zheng, M. Wang, Q. Shen and J. Huang, "Object-Scale Adaptive Convolutional Neural Networks for High-Spatial Resolution Remote Sensing Image Classification," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 283-299, Dec. 2021.

[34] Jie H , Li S , Gang S , et al, "Squeeze-and-Excitation Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011- 2023, April 2019.

[35] Z. Tian, W. Wang, R. Zhan, Z. He, J. Zhang and Z. Zhuang, "Cascaded Detection Framework Based on a Novel Backbone Network and Feature Fusion," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 9, pp. 3480-3491, Sept. 2019.

[36] Z. Gao, H. Ji, T. Mei, B. Ramesh and X. Liu, "EOVNet: Earth-Observation Image-Based Vehicle Detection Network," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 9, pp. 3552-3561, Sept. 2019.

[37] G. Xia et al., "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, pp. 3974-3983, June 2018.

[38] 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, no. 12, pp. 119–132, Dec 2014.

[39] Redmon J , Farhadi A, "YOLO9000: Better, Faster, Stronger," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, pp.6517-6525, July 2017.

[40] Dai J , Li Y , He K , et al. "R-FCN: Object Detection via Region-based Fully Convolutional Networks[C]// Advances in Neural Information Processing Systems". Curran Associates Inc. 2016.

[41] Redmon J and Farhadi A, "Yolov3: An incremen tal improvement," arXiv preprint arXiv:1804.02767, 2018.

[42] Fu C Y , Liu W , Ranga A , et al. DSSD : Deconvolutional Single Shot Detector[J]. 2017.

[43] Acatay O , Sommer L , Schumann A , et al. "Comprehensive Evaluation of Deep Learning based Detection Methods for Vehicle Detection in Aerial Imagery"，IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018.

[44] Xingjia Pan, Y uqiang Ren, Kekai Sheng, Weiming Dong, Haolei Y uan, Xiaowei Guo, Chongyang Ma and Changsheng Xu, "Dynamic Refinement Network for Oriented and Densely Packed Object Detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, pp. 11204-11213, June 2020.

[45] Xue Yang, Junchi Yan, Ziming Feng and Tao He, "R3det:Refined single-stage detector with feature refinement for rotating object," Proc. IEEE AAAI., Virtual Event, pp.11207–11216, February 2021.

[46] Kun Fu, Zhonghan Chang, Yue Zhang and Xian Sun, "Pointbased estimator for arbitrary-oriented object detection inaerial images," IEEE Trans. Geosci. Remote Sens., pp.1-18, Sep 2020.

[47] Gong C, Zhou P and Han J, "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, Sep 2016.

[48] 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, no. 1, pp. 119–132, April 2014.

[49] Kong T , Yao A , Chen Y , et al, "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, pp. 845-853, June 2016.

[50] Lee Y , Hwang J W , Lee S , et al. "An Energy and GPU-Computation Efficient Backbone Network for Real-Time Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2019.

[51] Ma N , Zhang X , Zheng H T , et al. "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," European Conference on Computer Vision. Springer, Cham, 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3079968, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

[52] Z. Li and F. Zhou. Fssd: Feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960, 2017.

[53] J. Jeong, H. Park, and N. Kwak. Enhancement of ssd by concatenating feature maps for object detection. In BMVC, 2017.

Qiong Ran received her Ph.D. degrees from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, China, in 2009. She has published over 10 papers in China and abroad. She is currently with the College of Information Science and Technology at Beijing University of Chemical Technology, Beijing, China. Her research interests include image acquisition, image processing, hyperspectral image analysis and applications.

Qing Wang received his bachelor degree from Qingdao Institute of Technology, China, in 2017. He is currently persuing his master's degree in College of Information Science and Technology at Beijing University of Chemical Technology, Beijing, China.

Boya Zhao received his B.Sc. degree from the School of Electrical Engineering and Information, Hebei University of Technology, Tianjin, China, in 2013 and Ph.D. degree in School of Electrical and Information Engineering, Beijing Institute of Technology, Beijing, China, in 2019. He is currently an assistant professor with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interest includes the object detection in complex background and on-board real-time information processing.

Yuanfeng Wu received the B.S. and M.S. degrees in computer science from China University of Mining and Technology, Beijing, China, and the Ph.D. degree in cartography and geographical information system from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2010. He is currently an Associate Professor with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include the development of onboard real-time algorithms, high-performance computing implementation and computer software in hyperspectral image processing.

Shengliang Pu is with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: pusl@aircas.ac.cn)

Shengliang Pu received the B.S. degree in Geodetic Engineering from the School of Geodesy and Geomatics, Wuhan University, China, in 2009; the M.S. degree in Software Engineering from the College of Computer Science, Inner Mongolia University, China, in 2013; and the Ph.D. degree in Photogrammetry and Remote Sensing from the School of Geodesy and Geomatics, Wuhan University, in 2019. He is currently a postdoc researcher with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China; and a lecturer with the Faculty of Geomatics, East China University of Technology, Nanchang, China. He has a wide variety of research interests in Remote Sensing. Currently, his research is mainly focused on hyperspectral remote sensing image processing.

Zijin Li received the B.S degree from the School of Earth and Space Sciences, Peking University, Beijing, China, in 2021.She

is working toward the M.S. degree at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interest includes hyperspectral image processing and object detection.