Neurocomputing 451 (2021) 67–80

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery

Luchang Chen, Chunsheng Liu *, Faliang Chang, Shuang Li, Zhaoying Nie

School of Control Science and Engineering, Shandong University, Jinan 250061, China

ARTICLE INFO

Article history:
Received 6 November 2020
Revised 15 March 2021
Accepted 2 April 2021
Available online 6 April 2021
Communicated by Zidong Wang

Keywords:
Object detection in aerial images
Deep neural network
Attention network
Training strategy

ABSTRACT

Compared with the classic object detection problem, detecting objects in aerial images has some special challenges including huge orientation variations, complicated and large background, and wide multi-scale distribution. Considering these three challenges together, we propose a novel arbitrary-oriented object detection framework consisting of three main parts. Firstly, the *Cascading Attention Network* (CA-Net) composed of a patching self-attention module and a supervised spatial attention module is proposed for enhancing the feature representations from objects of interest and suppressing the background noises in *Feature Pyramid Network* (FPN) from coarse to fine. Then, the *Adaptive Feature Concatenate Network* (AFC-Net) is proposed to adaptively stack the feature maps pooled from all FPN levels as well as the global semantic features, for dealing with the multi-scale change of objects. Lastly, the *OBB Multi-Definition and Selection Strategy* (OBB-MDS-Strategy) is proposed to regress rotated bounding boxes more smoothly and detect oriented objects more accurately in the training process. Our experiments are conducted on two common and challenging aerial datasets, i.e., DOTA and HRSC2016. Experiments results show that the proposed method has superior performances in multi-orientated objects detection compared with the representative methods.

1. Introduction

With the rapid development of remote sensing technology and aerial shooting technology, more and more high-resolution aerial images can be easily acquired for aerial image analysis. Multi-class object detection plays an important role in automatic analysis of aerial images, and it also becomes critically important for a wide range of applications such as intelligent monitoring, urban planning, and precision agriculture [1]. Compared with the objects detection methods in natural scenes, locating the specified objects and recognizing their categories in aerial images have some special challenging problems such as huge orientation variations (Fig. 1(a)), complicated and large background (Fig. 1(b)) and wide multi-scale distribution (Fig. 1(c) and (d)).

These three factors make it challenging to detect objects in aerial images. There are some previous methods that are aimed to address these three problems.

For the problem of huge orientation variations, *horizontal bounding box* (HBB) based methods and *oriented bounding box* (OBB) based methods have different performance. Most traditional detection methods are HBB based methods, which has achieved considerable progress in some fields such as person detection, vehicle detection, and general object detection; yet HBB methods usually have some problems when detecting objects in aerial images which have a variety of object orientations as shown in Fig. 1(a). With large redundant regions, the HBB methods usually cannot locate objects accurately. Many OBB methods have been proposed for multi-oriented object detection, such as *Rotational Region CNN for Orientation Robust Scene Text Detection* (R2CNN) [2], *Rotated region based CNN for ship detection* (RRCNN) [3], *Context-Aware Detection Network for Objects in Remote Sensing Imagery* (CAD-Net) [4], etc. These OBB methods have more accurate bounding boxes than HBB based methods. Yet, these methods directly regress one extra angle parameter theta of an OBB, which may cause angle mutation problem; according to OpenCV's angle definition, the angle may have a huge change near -90°, which may result in poor detection performance. Yang et al. [5] propose an IoU-smooth L1 loss to solve the angle mutation problem. But the IoU-smooth L1 loss only suppresses the process of OBB regression and cannot solve the problem fundamentally.

Complicated and large background in aerial images is also a challenging problem for detecting objects especially with small-size. Many attention-based methods have been proposed to guide

* Corresponding author.
E-mail address: liuchunsheng@sdu.edu.cn (C. Liu).

https://doi.org/10.1016/j.neucom.2021.04.011
0925-2312/

Fig. 1. Some aerial images in the DOTA dataset. (a) The direction of objects in aerial images may be arbitrary. HBB contains more redundant regions, while OBB is with less redundant regions. (b) Aerial images usually contain complicated and large background. (c) The scales of the objects are widely distributed, such as bridges of different sizes in different aerial images. (d) The scales of objects shot in the same image may also vary greatly.

the network to pay more attention to object areas and ignore the background noises. *Multi-Layer Attention Network* (MANet) [6] is proposed to make the network to focus more on the location of targets by combining a position attention block and a channel attention block. Li et al. [7] propose a *Feature-Attentioned Feature Pyramid Network* (FA-FPN) with a dot-product attention module to capture the foreground information and restrain the background. However, these unsupervised attention based methods cannot achieve the special purpose of learning to suppress the background information.

For the problem of wide multi-scale distribution, many previous methods have discovered the effect of constructing a multi-layer network. *Single Shot MultiBox Detector* (SSD) [8] performs detection on multi-scale feature maps of different strides, which may lose the contextual information; *Feature Pyramid Network* (FPN)[9] is a classic network to deal with the scale problem. It uses a feature pyramid with a top-down pathway to combine the strong semantic features on the top layer and the high-resolution information on the bottom layer. FPN-based detectors detect multi-scale objects by assigning Region-of-Interests to different feature levels of FPN according to scales and then extract features. Yet, this assigning strategy has some problems: (1) each RoI is just assigned to a single-level of FPN and extracts features, which is not enough for detection; (2) some RoIs may be assigned to an inappropriate layer and lose important information. These problems are more obvious for detecting objects with unbalanced sizes in remote sensing images.

In this study, we consider three main problems together and propose a novel rotation detection framework called *Adaptive Multi-Level Feature Fusion and Attention-Based Network* (AMFFA-Net), which consists of three main parts including the *Cascading Attention Network* (CA-Net), the *Adaptive Feature Concatenate Network* (AFC-Net) and the *OBB Multi-Definition and Selection Strategy* (OBB-MDS-Strategy). Firstly, the CA-Net is proposed with a novel cascading attention structure for the purpose of suppressing the background noises in FPN from coarse to fine. The cascading attention structure has two parts including the *Patching Self-attention Module* (PSA-Module) and the *Supervised Position Attention module* (SPA-Module); the PSA-Module is utilized to enhance the feature representations from objects of interest; the SPA-Module is introduced to specifically filter out background noises by a supervised learning method. Secondly, in order to overcome the problems caused by assigning strategy in FPN, the AFC-Net is proposed to adaptively merge the features from different pyramid levels for each region proposal and obtain multi-level features, dealing with the wide multi-scale distribution of objects caused by the bird's wide view. Thirdly, to handle the huge orientation variation of objects, the OBB-MDS-Strategy is proposed, which can eliminate the angle mutation problem and detect oriented objects more accurately by redefining the original ground-truth OBB (GT OBB). All three parts constitute a novel rotation detection structure, which can largely overcome the challenges of huge orientation variations, complicated and large background, and wide multi-scale distribution in aerial images.

The experiments on two challenging aerial datasets including DOTA [10] and HRSC2016 [11] show that the proposed method has superior performance in multi-orientated object detection compared with the representative methods.

This work has made the following contributions:

- Aiming at guiding the network to focus more on object areas, the novel CA-Net with two attention blocks is proposed to suppress the background information in FPN.
- In order to better detect multi-scale objects, the novel AFC-Net is proposed to adaptively integrate features pooled from feature levels with different sizes.
- To handle the huge orientation variation of objects, the OBB-MDS-Strategy is proposed to eliminate the influence of angle changes.

The remainder of this paper is organized as follows. In Section 2 we give a brief introduction about the related work. Section 3 presents different parts of the proposed structure. Experimental results including ablation study and comparison with other methods are presented in Section 4, and Section 5 gives our conclusions and future work.

2. Related works

2.1. Generic horizontal object detector

Due to the development of convolutional neural networks based on deep learning, the performance of object detection has been greatly improved. Many classic object detection methods aim to represent an object with a horizontal bounding box has been proposed, and can be roughly divided into two-stage and one-stage methods. Two-stage detection methods are mainly region-based, such as R-CNN [12] and its variances Fast R-CNN [13], Faster R-CNN [14], and R-FCN [15], which generate many regions of interests from the input image through the Selective Search or region proposal network in the first stage and then use these region features to predict the category of object and refine the horizontal bounding box. *Feature Pyramid Network* (FPN) [9] is proposed to deal with large object scale variation in natural images by building a feature pyramid which generates feature maps of different scales from different layers. Cascade R-CNN [16] performs multiple refinement in the second stage of

region-based detector, which achieves more accurate predictions of object classification and object location.

Single-stage detection methods achieve a faster speed but not as accurate as a two-stage detector. You Only Look Once (YOLO) [17] uses the entire image as the input of the network, and directly performs classification and regression on the output layer. SSD [8] sets prior anchor boxes on multi-scale feature maps and makes classification and regression based on these anchors. Retina-net [18] proposes a novel Focal loss to solve the imbalance problem of positive and negative samples in one-stage detectors.

2.2. Oriented object detector

Most of the objects in the remote sensing image and text in the natural scene have arbitrary-orientation. Using oriented bounding boxes to represent them is more accurate than horizontal bounding boxes. For scene text detection, Ma et al. [3] present a rotated RPN (RRPN) network to generate rotated proposals and uses Rotation Region-of-Interest (RRoI) pooling layer in the second stage of Faster R-CNN to regress the oriented bounding box. Jiang et al. [2] present a Rotational Region CNN (R2CNN) method uses ROI layer with different sizes to accommodate texts of different sizes and aspect ratios. Textboxes++ [19] directly regresses 8 vertices of OBB on SSD. RRD [20] improves the Textboxes++ by extracting rotation-invariant and rotation-sensitive features to do classification and regression respectively. Zhao et al. [21] propose an Elite Loss in segmentation based text detection networks to pay more attention to the on-stroke pixels. For aerial images object detection, DRBox [22] adds prior anchor boxes with multiple angles on the basis of SSD [8] to regress the oriented bounding boxes. Liu et al. [3] exploit RRPN and multi-tasks NMS for rotated ship detection. R-DFPN [23] solves the narrow width problem of the ship by proposing a Dense Feature Pyramid Network. Ding et al. [24] propose an RoI transformer module which transforms the horizontal proposals to rotated one and regress the oriented bounding box. Azimi et al. [25] employ an image-cascade network and a feature pyramid network with kernels of different size to extract multi-scale features. Wang et al. [26] propose a rotation detector named SegmRDet, which leverage a segmentation task to generate the rotated bounding box by identifying a minimum-sized bounding box from the predicted foreground pixels. Fu et al. [27] propose a novel point-based estimator that operates a fully convolutional network to localize an object as a set of points, thereby explicitly utilizing the spatial information; moreover, oriented object detection is separated by sequential operating instance localization and object recognition, eliminating the problem of feature construction. Anchor-free detection method [28–31] has been widely used in many fields and is also used in aerial images to deal with the problem of large differences in object aspect ratios. Both [32,33] detect orientated objects in per-pixel prediction fashion, regress the vertices or angles of the bounding box for the positive sample pixels. Chen et al. [34] propose a target heat-map network (THNet) to predict the heat-map of each class and identify their positions by a thresholding based location method.

Most of the aforementioned anchor-based methods, regressing OBBs from rough horizontal proposals, do not consider the angle mutation problem, which may lower the performance of rotation detector.

2.3. Multi-scale detector

Scales of objects in natural images are not consistent, using features from single level to represent them is inappropriate. Thus several methods have been raised to come up with multi-scale object detection by using features from multi layers of CNN. Liu et al. [8] propose SSD which uses multi-scale feature maps from different layers to do the detection task. FPN [9] fuses multi-scale features via the top-down pathway and lateral connection to produce more semantic information. Inspired by FPN, some works [7,25,23] enhance the semantic representation by re-constructing pyramid network. Libra-rcnn [35] uses a balanced feature to enhance original features by combining each resolution in the pyramid.

These FPN-based detectors always select a single level from feature pyramid to pool features, which cannot satisfy the detection task in aerial images due to the wide scale distribution of objects. And the strategy of RoI assigned to FPN may result in losing appropriate scale information.

2.4. Attention mechanism

Visual attention mechanism is a useful way to reject large irrelevant background information and focus on related objects. Jaderberg et al. propose *Spatial Transformer Networks* (STN) [36], performing spatial transformation on the spatial domain to extract key information. Wang et al. [37] propose Non-local Neural Network, which correlates different locations of a pixel to calculate long-range dependencies with other pixels. Zhang et al. [38] propose Self-Attention module, which allows attention-driven, long-range dependency modeling for image generation tasks. Li et al. [7] insert the self-attention module [38] as lateral connections in FPN to improve the visual representation and suppress the complex background noises of aerial images. Zhang et al. [4] propose a Spatial-and-Scale-Aware Attention Module to obtain an attention-modulated feature map which focuses on informative regions of aerial images.

Aerial images are always shot with a large area of view, which results in images containing many elements, especially background elements. Both [37,38] model the relationship of pixels that far away in very large aerial images, which is a redundant calculation and may bring in disturbing information. Spatial attention methods like [4] are only updated through the loss backpropagation of label and final predict value of the whole network, which has difficulty to satisfy the purpose of suppressing background noises. Our proposed CA-Net captures the pairwise long-range dependencies in local areas of huge aerial images and introduces mask information to supervise the network to capture the foreground information. Experiments in Section 4.4 demonstrate the effectiveness of CA-Net.

3. Methods

Compared with the classic object detection problem, detecting objects in aerial images has some special challenges including huge orientation variations, complicated and large background, and wide multi-scale distribution. To overcome these difficulties, we propose a novel network named *Adaptive Multi-Level Feature Fusion and Attention-Based Network* (AMFFA-Net). As illustrated in Fig. 2, the proposed AMFFA-Net is a two-stage OBB detector, which consists of three main parts including the *Cascading Attention Network* (CA-Net), the *Adaptive Feature Concatenate Network* (AFC-Net) and the *OBB Multi-Definition and Selection Strategy* (OBB-MDS-Strategy). First of all, in order to make object features more prominent in the complex background, the CA-Net is proposed to restrain redundant noises and capture more useful foreground information. Then, the AFC-Net is designed to adaptively aggregate multi-level features to take full advantage of all level features. In the second stage of this detector, the OBB-MDS-Strategy is proposed to regress the rotated bounding box smoothly. Finally, the detection results of objects with arbitrary directions can be

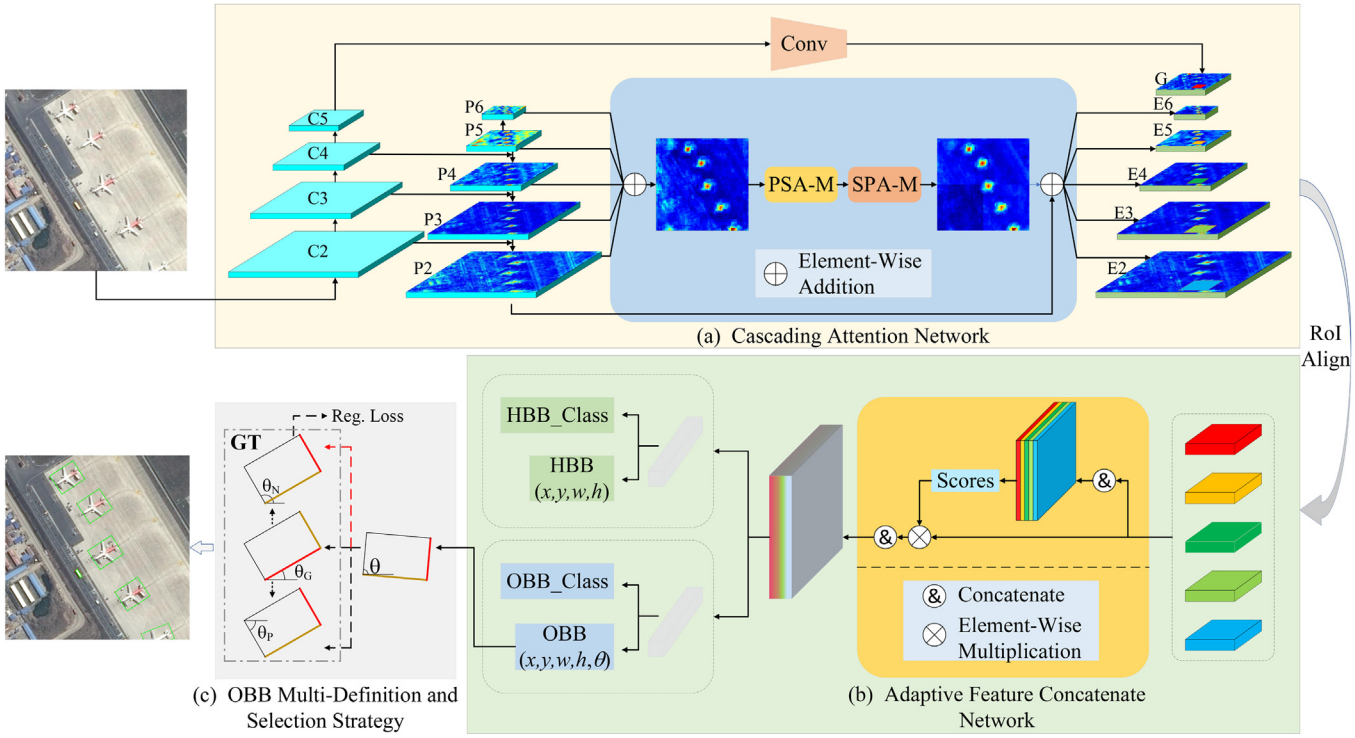

Fig. 2. The overall framework of the proposed AMFFA-Net. (a) shows the *Cascading Attention Network* (CA-Net). The CA-Net enhances the features of the original FPN by suppressing the background information. (b) is the *Adaptive Feature Concatenate Network* (AFC-Net). The AFC-Net adaptively concatenates the pooled features from all levels of FPN by learning spatial weights to take advantage of multi-level features. (c) shows the *OBB Multi-Definition and Selection Strategy* (OBB-MDS-Strategy). The red line and the orange line represent the width and height of Ground Truth OBB (GT OBB) respectively. The OBB-MDS-Strategy first reverses the angle and changes the height and width of the original GT OBB to obtain another two OBB definitions. Then the OBB which has the same definition (i.e. the top GT OBB in (c)) as the predicted OBB will be selected as the true GT OBB. The true GT OBB is used to calculate the regression loss.

obtained. These three main parts will be described in detail in the following subsections.

3.1. Cascading attention network

Aerial imagery has a wide perspective with complex natural backgrounds, and may interfere with region proposal network (RPN) [14] to generate many false proposals and miss correct detections. Some visual attention methods have been proved useful dealing with this problem in some computer vision tasks [39–43], but these attention modules are not directly designed for aerial images.

In this study, we propose the *Cascading Attention Network* (CA-Net) for aerial images, with the purpose of making FPN focus more on information-rich object areas and help the RPN generate more accurate region proposals. The CA-Net has two parts including the patching self-attention block and the supervised spatial attention block. Inspired by self-attention module [38], the proposed patching self-attention block divides the feature map outputted by the feature extraction network into four patches and establishes rich context relationships in each patch to enhance the feature representations from objects of interest, which can largely reduce the great computation requirement of [38]. At the same time, the calculation between two pixels that are not closely related in large aerial images is avoided. Features refined by the patching self-attention block are discriminative and still have some background noises, and the supervised spatial attention block aims to suppress these background noises by learning an attention map.

The whole architecture of CA-Net is shown in Fig. 3. The outputs of the residual block in each stage of ResNet [44] are denoted as $\{C2, C3, C4, C5\}$. Feature pyramids $\{P2, P3, P4, P5\}$ are constructed by combining features of two adjacent levels of *Ci*. Firstly, the features *Pi* are rescaled to the size of *P*4 by a bilinear interpolation and integrate them by an element-wise addition operation to obtain the fused feature F^f, with the purpose of balancing the low-level spatial information and high-level semantic information.

$$F^f = P4 + \sum_{i=2}^{3} Rescale(Pi) + \sum_{i=5}^{6} Rescale(Pi) \tag{1}$$

Secondly, the fused feature F^f is fed into the first part of CA-Net, named Patching Self-Attention Module (PSA-Module); the PSA-Module module is designed to eliminate the aliasing effect caused by interpolation in feature fusion process and to establish rich context relationships on local features. The number of multiplications in traditional self-attention is $(2N^2C)$. Our PSA-Module splits the whole feature map $F^f \in \mathbb{R}^{H\times W\times C}$ into four patches $F_i^f \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$, and just need to calculate correlation of each pixel with other pixels within the belonged patch. In this way, the number of multiplications is reduced to $(\frac{N^2C}{2})$ and the calculations between irrelevant pixels are also reduced, which is more suitable for large aerial images. After reducing the channel from C to $C/8$ by a 1×1 Conv, each of these patching feature maps is reshaped to $A \in \mathbb{R}^{C/8\times N}$ and $B \in \mathbb{R}^{N\times C/8}$, where $N = 32 \times 32$ (A quarter of the size of *P*4) in this study. Next a matrix multiplication is performed between *A* and *B* followed by a softmax function to obtain the attention map for each patching feature map. By multiplying the attention map with the reshaped original feature *C*, we get features weighted by local pixel attention.

$$F_i^m = F_i^f + C \cdot Softmax(A \cdot B),\ i = 0,\ 1,\ 2,\ 3 \tag{2}$$

where, F_i^m is the patch of F^m, F^m is the middle refined feature outputted by PSA-Module.

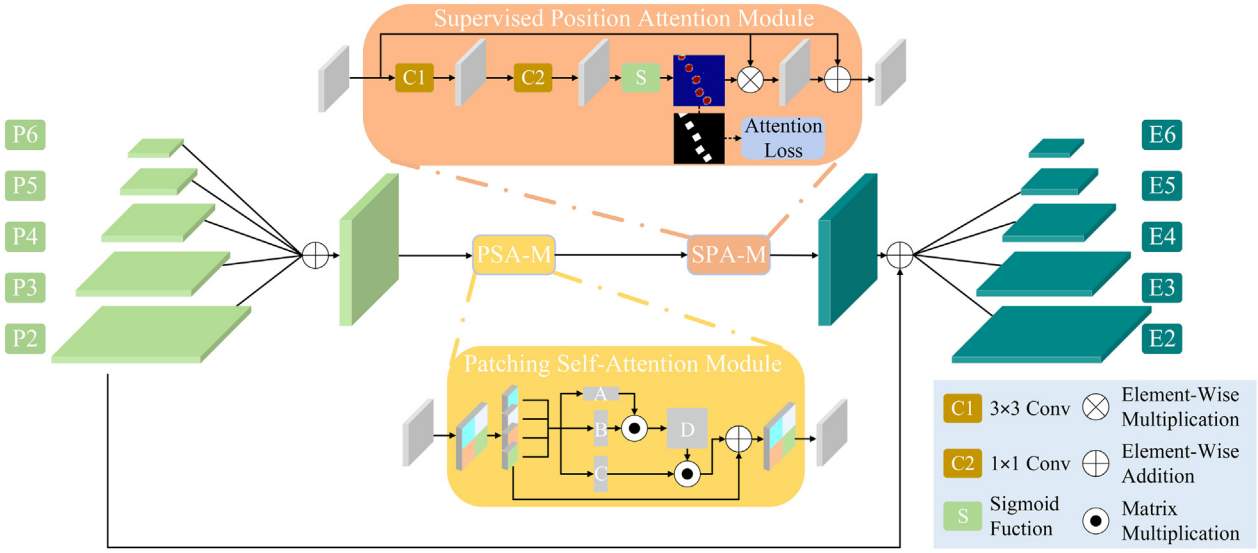

Fig. 3. Overview of the proposed Cascading Attention Network (CA-Net).

Thirdly, the middle refined feature F^m from PSA-Module is fed into the second part of the cascading attention module, named Supervised Position Attention Module (SPA-Module); the SPA-Module introduces the mask guided mechanism, guiding the attention map to filter out background information, and also helps to retain some contextual information.

More specifically, the refined features pass through a 3×3 Conv to extract contextual information, and then a one-channel feature map is learned through a 1×1 convolution operation. The feature map represents the score of the foreground. After that a sigmoid operation is performed on the score map and multiplied with refined features, and then a new refined feature map is obtained. For supervision, the cross-entropy loss of the score map and the binary mask is used to define the attention loss. Because there is usually no precise mask annotation in the aerial images, we generate the ground-truth mask by filling the ground-truth oriented bounding box as 1 and other position as 0 as shown in Fig. 3. All the ground-truth oriented bounding boxes in one image are filled as 1 regardless of classes, with the purpose of highlighting foreground regions (like saliency object detection method [45]). These generated samples are the ground-truth masks in the SPA-Module. The refined feature map is determined as,

$$M = \sigma(\phi(F^m)) \tag{3}$$

$$F^r = F^m + M * F^m \tag{4}$$

where $\sigma(\cdot)$ is a sigmoid function, $\phi(\cdot)$ is the operation of Convolutional layers. M is the attention map, F^r is the refined feature map.

Considering that the refined feature map has more useful information and ignores irrelevant information, we use it to improve the original FPN feature. Notice that the size of the refined feature map is the same of $P4$, so we rescale the refined feature map to the same size as $\{P2, P3, P4, P5, P6\}$ respectively which are donated as $\{R2, R3, R4, R5, R6\}$. Then Ri are added to corresponding Pi to obtain the final enhanced FPN features $Ei = \{E2, E3, E4, E5, E6\}$.

$$Ei = Pi + Ri \tag{5}$$

The comparison between feature maps $\{P2 - P6\}$ and $\{E2 - E6\}$ is shown in Fig. 9. From Fig. 9, it can be seen that the feature maps $\{E2 - E6\}$ have less background noises and focus more on objects than $\{P2 - P6\}$.

3.2. Adaptive feature concatenate network

The FPN structure extracts features from $\{P2 - P5\}$ ($P6$ is not used in the detection head) for multi-size objects according to the size of the proposal. Generally, the proposals of small objects are assigned to low-level features, the proposals of big objects are assigned to high-level features. The assigning principle is represented as Eq. (6).

$$k = \lfloor k_0 + \log_2(\sqrt{wh}/224) \rfloor \tag{6}$$

where k presents the assigned level, $\lfloor \cdot \rfloor$ is the rounding down operation, 224 is the standard ImageNet pre-training size, k_0=4 means that the proposal of size 224×224 should be assigned to $P4$, w and h represent the width and height of the proposal.

For objects with obvious scale differences in remote sensing images, this assigning principle may cause two problems: features of each proposal are only extracted from the single-level feature map of FPN, which is not enough to satisfy the multi-scale detection task; RoIs with similar size may be assigned to the different feature levels of FPN, which causes some of them to ignore important information in the proper feature level.

To solve these problems, we propose the *Adaptive Feature Concatenate Network* (AFC-Net) to fuse pooled features of each ROI from all feature levels and the global semantic features extracted by the backbone network. At the same time, considering that the importance of pooled features from each level is not the same, AFC-Net measures the importance of these pooled features and integrates them according to the learned weights. The details are presented as follows.

First, each proposal is assigned to all pyramid levels, i.e. $\{E2 - E5\}$, to get the pooled features (denoted as L_1, L_2, L_3, L_4) by RoI align operation [46]. Meanwhile, the same operation is utilized to get the pooled features (denoted as G) from the global semantic features. The global semantic features come from the final feature maps of the ResNet backbone, and the channels have been reduced by a convolutional layer. Then the AFC-Net measures the importance of these pooled features and adaptively concatenate them by the learned spatial weights. The structure of AFC-Net is shown in Fig. 4. Formally, given five pooled features L_1, L_2, L_3, L_4, G, we first use five 1×1 convolutional layers to reduce their channels respectively and then aggregate them via concatenation.

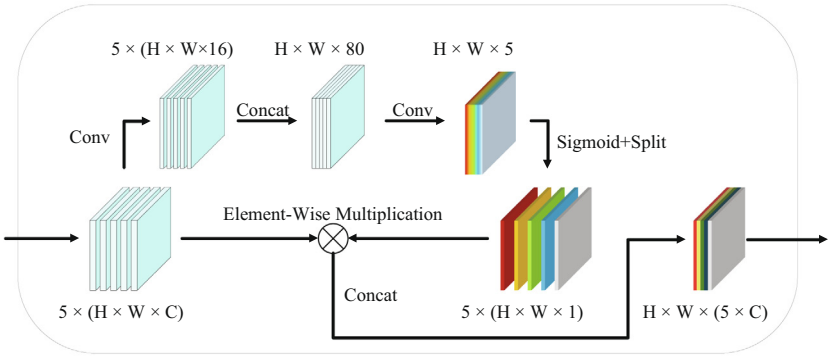

Fig. 4. The network structure diagram of the Adaptive Feature Concatenate Network (AFC-Net).

$$L_c = concat(W_1^1(L_1),\ W_1^2(L_2),\ W_1^3(L_3),\ W_1^4(L_4),\ W_1^5(G)) \tag{7}$$

where, $W_1^i \in \mathbb{R}^{C\times 16}$ is parameterized as a 1×1 convolutional layer, $i = 1, 2, 3, 4, 5$; ,L_c is the aggregated feature map and $L_c \in \mathbb{R}^{H\times W\times 80}$. Then another 1×1 convolutional layer and a sigmoid function are used to generate the spatial weights map.

$$S = \gamma(W_2(L_c)) \tag{8}$$

where $\gamma(\cdot)$ denotes the sigmoid function, $W_2 \in \mathbb{R}^{80\times 5}$ is parameterized as another 1×1 convolutional layer.

Then the weights map S is split for each pooled feature donated as $S_i \in \mathbb{R}^{H\times W\times 1}$, $i = 1, 2, 3, 4, 5$. After that, we multiply the spatial weights S_i and the pooled features $\{L_1, L_2, L_3, L_4, G\}$ by element-wise multiplication and concatenate them to obtain the integrated region feature F_N.

$$F_N = concat(S_1 * L_1,\ S_2 * L_2,\ S_3 * L_3,\ S_4 * L_4,\ S_5 * G) \tag{9}$$

The F_N is sent to detection head to generate final results.

3.3. OBB multi-definition and selection strategy

3.3.1. Oriented bounding box

In the detection head, the regional feature F_N extracted by AFC-Net passes through two fully connected layers to generate the predicted OBB with five parameters (x, y, w, h, θ) and predict its corresponding category. The representation of the five parameters is the same as OpenCV, as shown in Fig. 5:

- Among the four sides of a rotated quadrilateral, the side at the lower right corner is defined as width w, and its adjacent side is defined as height h. For a horizontal quadrilateral, the side perpendicular to the x-axis is defined as w and its adjacent side is defined as h.

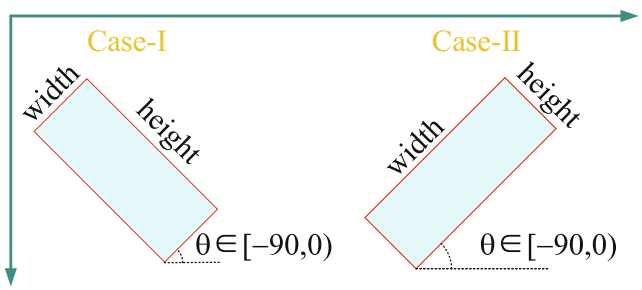

Fig. 5. The representation of OBBs. Two OBBs in different cases have different representation.

- The rotation angle θ of a quadrilateral is defined as the angle between the width side and the positive direction of the x-axis which ranges from -90° to 0°.
- (x, y) is the center point coordinates of the quadrilateral.
- In order to facilitate the following analysis, when the short side of an OBB is w, we call this OBB Case-I; when the short side of an OBB is h, we call this OBB Case-II.

3.3.2. Similar feature error

When GT OBB of an instance, with angle close to -90° or 0°, belongs to Case-I (or Case-II), the predicted OBB of this instance may be Case-II (or Case-I) because of the similar feature error shown in Fig. 6. As shown in Fig. 6, the regional feature in Fig. 6 (b) is similar to the regional feature in Fig. 6(c), resulting in predicting an incorrect but appropriate OBB (IA-OBB) which belongs to a different case of GT OBB. In Fig. 6, ideally, IA-OBB only needs to rotate clockwise to become a correct OBB, and the loss between it and the GT OBB is relatively small. But according to the OBB representation shown in Fig. 5, IA-OBB needs to significantly scale its width and height and reverse the angle to become a correct OBB; in this situation, the loss between IA-OBB and GT OBB is relatively large, resulting in difficulty in box regression.

3.3.3. Selected true OBB

The instability of the network is caused by different OBB expressions under similar features. The novel *OBB Multi-Definition and Selection Strategy* (OBB-MDS-Strategy) for GT OBB is designed to deal with this problem. The process of OBB-MDS-Strategy is shown in Fig. 7. Specifically, before calculating the loss between the predicted OBB and the corresponding GT OBB in the detection head, another two expressions of this GT OBB are generated. This set of OBBs is,

$$OBB_{set} = \begin{cases} OBB_1 = (x,\ y,\ w_1 = w,\ h_1 = h,\ \theta_1 = \theta_G) \\ OBB_2 = (x,\ y,\ w_2 = h,\ h_2 = w,\ \theta_2 = \theta_P) \\ OBB_3 = (x,\ y,\ w_3 = h,\ h_3 = w,\ \theta_3 = \theta_N) \end{cases} \tag{10}$$

where, $\theta_P = \theta_G + 90^\circ$, $\theta_N = \theta_G - 90^\circ$. Notice that two extra definitions are added to the original OBB: one increases the original angle by 90°, and the other reduces the original angle by 90°, and both of them exchange the height and width. The goal is to select a true OBB from formula (10) to minimize the error between the predicted OBB and the GT OBB. Thus true OBB is obtained by:

$$\begin{aligned} &True\ OBB = OBB_{set}[j], \\ &j = arg\ min[loss(OBB_{set}[i],\ Predicted\ OBB),\ i = 0,\ 1,\ 2] \end{aligned} \tag{11}$$

where *loss* refers to Smooth_{L_1} loss. Finally the true OBB with five parameters (x, y, w, h, θ) is preserved as feedback. For example, the definition of yellow OBB in Fig. 6(a) is changed to (0, 0, 10,

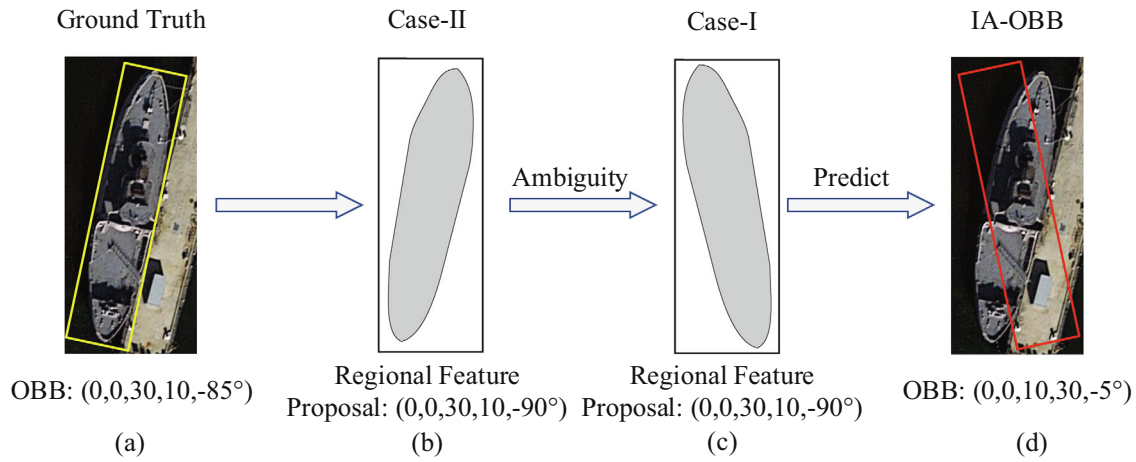

Fig. 6. The explanation of similar feature error. (a) A ship with yellow Ground Truth OBB: (0, 0, 30, 10, −85°). (b) The corresponding region feature of the ship, whose coordinates of proposal (black rectangle) is (0, 0, 30, 10, −90°). (c) The similar feature, whose coordinates of proposal is (0, 0, 30, 10, 90°), too. (d) The incorrect but appropriate predicted OBB (IA-OBB); red rectangle indicates the IA-OBB. Notice that the GT OBB and IA-OBB are in different cases: (0, 0, 30, 10, −85°) versus (0, 0, 10, 30, −5°).

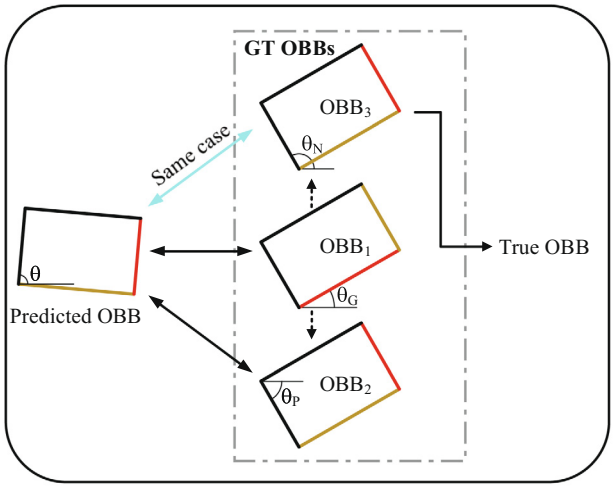

Fig. 7. Example of OBB Multi-Definition and Selection Strategy. The red slide and the orange slide represent the width and height of OBB respectively. OBB_1 is the original GT OBB, OBB_2 and OBB_3 are converted from the original GT OBB. According to the rule specified in Section 3.3.1, OBB_3 and OBB_2 have the same case as predicted OBB. But the angle of OBB_3 is close to the angle of predicted OBB, so the loss between OBB_3 and the predicted OBB is relatively small, means that the regress process of predicted OBB will be easier and smoother.

30, 5°) by OBB-MDS-Strategy. In this way, the red predicted OBB in Fig. 6(d) just needs to rotate 10 degrees clockwise to be the correct OBB.

OBB-MDS-Strategy does not involve some complex network structures, but only improves the definition of OBB, which can be used by any horizontal anchor-based method to improve performance.

3.4. Learning of oriented bounding box

The loss function for oriented bounding box prediction takes the form of multi-task loss, including regression and classification of HBBs and OBBs, pix-wise classification for attention mask learning, defined as,

$$L(p,\ c,\ c^*,\ v) = \lambda_1 L_{cls}(p,\ c) + \lambda_2 L_{reg}(v*,\ v) + \lambda_3 L_{seg}(A^*,\ A) \tag{12}$$

where, c is the class label of a proposal, and p is the output probability. $v = (v_x,\ v_y,\ v_w,\ v_h,\ v_\theta)$ represents the predicted offset vector for class label c, and $v^* = (v_x^*,\ v_y^*,\ v_w^*,\ v_h^*,\ v_\theta^*)$ represents the target vector. A^* and A represents the ground-truth mask and the mask learned by network respectively. The balancing parameters λ_1, λ_2 and λ_3 are fixed as $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in this study.

L_{cls} is the loss of the classification branch,

$$L_{cls}(p,\ c) = -log(p) \tag{13}$$

$L_{reg}(v^*,\ v)$ is designed for the oriented bounding box regression branch,

$$L_{reg}(v^*,\ v) = \sum_{i \in \{x,\ y,\ w,\ h,\ \theta\}} Smooth_{L_1}(v_i^* - v_i) \tag{14}$$

and,

$$Smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if\ |x| < 1, \\ |x| - 0.5, & otherwise, \end{cases} \tag{15}$$

and, v and v^* are defined as,

$$\begin{aligned} v_x = (x - x_a)/w_a,\ v_y = (y - y_a)/h_a \\ v_w = log(w/w_a),\ v_h = log(h/h_a) \\ v_\theta = \theta - \theta_a \end{aligned} \tag{16}$$

$$\begin{aligned} v_x^* = (x^* - x_a)/w_a,\ v_y^* = (y^* - y_a)/h_a \\ v_w^* = log(w^*/w_a),\ v_h^* = log(h^*/h_a) \\ v_\theta^* = \theta^* - \theta_a \end{aligned} \tag{17}$$

where, variables x, x^*, x_a refer to GT OBB (i.e. true OBB obtained by OBB-MDS-Strategy), predicted OBB and anchor box respectively (similarly for y, w, h, θ). All the anchor boxes are horizontal, so θ_a always equal to −90°.

For the attention branch, $L_{seg}(A^*,\ A)$ is a pixel-wise sigmoid cross entropy loss,

$$L_{seg}(A^*,\ A) = \frac{16}{W * H} \sum_{i}^{W/4} \sum_{j}^{H/4} (-A^*(i,\ j)\ \ln(A(i,\ j))) \tag{18}$$

where, $A^*(i,\ j)$ and $A(i,\ j)$ refer to the ground-truth mask label and the predicted attention map value at each position $(i,\ j)$.

4. Experiments and results

We evaluate the proposed network on two data sets: DOTA [10] and HRSC2016 [11], for object detection in remote sensing images.

4.1. Datasets

4.1.1. DOTA

DOTA [10] dataset is the largest dataset for object detection in aerial images with inclined bounding box annotations. It contains 2,806 large size images and 188, 282 instances of 15 object categories: plane, baseball diamond (BD), bridge, ground track field (GTF), small vehicle (SV), large vehicle (LV), ship, tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball field (SBF), roundabout (RA), harbor, swimming pool (SP) and helicopter (HC).

DOTA is split into training (1/2), validation (1/6), and test (1/3) sets. Only the ground truth annotations of training set and validation set are public, therefore we train and evaluate on training and validation set respectively. And the final prediction results on the test set are submitted to the official DOTA evaluation server when comparing our method with other published methods.

4.1.2. HRSC2016

HRSC [11] is a large dataset collected by Google Earth for ship detection. It includes 1,061 images and more than 2,000 instances with multiple directions. The image size ranges from 300×300 to 1500×900 pixels. The training set, validation set, and test set contain 436, 181, and 444 images, respectively.

4.2. Evaluation metrics

To evaluate the performance of our proposed method, we use the evaluation metrics same to the metrics used on PASCAL VOC. Specially, *Average Precision* (AP) and *mean Average Precision* (mAP) are adopted to evaluate the experimental results. They are defined as,

$$AP = \int_0^1 P(R)\, dR \tag{19}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{20}$$

where, P is precision and R is recall, $P(R)$ represents the curve composed of P and R, N indicates the number of class. Precision means the ratio of the number of correct positive predictions to the total number of positive predictions; Recall means the ratio of the number of true positive predictions to the total number of positive samples. The *Precision* and *Recall* are defined as,

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

and,

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

where, TP, FP and FN indicates the number of true positive predictions, the number of false positive predictions and the number of false negative predictions respectively.

4.3. Implementation details

All the experiments are implemented based on the deep learning framework of Pytorch [47]. The main configuration of our platform is with an Intel i7-6800K CPU, 32 GB DDR4, two NVIDIA 2080Ti graphics cards. The baseline is a rotation detection model extended from the FPN-based Faster R-CNN [14], and the OBB is regressed from coarse horizontal region proposal.

For the images in DOTA with large sizes, we split images in training dataset, validation dataset and testing dataset into the blocks of 1024×1024 with the overlap of 512 pixels using the official development kit. Both ResNet-50 and ResNet-101 are used as backbone when conducting experiments on DOTA. The model is trained for total 12 epochs with a batch-size 2. The initial learning rate is 0.005 and is decreased by 0.1 at epoch 8 and 11. Due to the limitation of the GPU memory, when the backbone is ResNet-101, the batch-size is set to 1 for each GPU, and the initial learning rate is set to 0.0025. For anchor details, the base anchor scale is set to 8, anchor ratios are set to [1/2, 1, 2]. The anchor is regarded as a positive sample when IoU between it and the axis-aligned bounding box enclosing GT OBB is greater than 0.7, while a negative sample is generated when IoU is less than 0.3. Besides, the two thresholds in the second stage are set to 0.5. When comparing our method with representative detectors, we apply data augmentation technologies. Multi-scale strategy is a common process for preprocessing DOTA data set, and most of the methods in comparative experiment have used it. For examples, the method in [24] resizes the images at two scales (1.0 and 0.5) for training and testing, and the method in [5] randomly scales the original images to [600×600, 800×800, $1{,}000 \times 1{,}000$, $1{,}200 \times 1{,}200$] and sends them to the network for training and testing; the method in [26] resizes the images at three scales (1.5, 1.0 and 0.5) for testing. Like [24], we resize the original images in training dataset and testing dataset at two scales (1.0 and 0.5) before splitting the images into patches. Furthermore, each image is randomly flipped with a probability of 0.5. In particular, all models for ablation studies are trained on the split training set of DOTA and evaluated on the split validation set of DOTA, because that the test set annotations of DOTA are not public and the number of submitting results to the official DOTA evaluation server is limited to twice a day. The final predictions are evaluated on the official test dataset for comparison with representative methods. When comparing with representative detectors, we first obtain predictions on the split test dataset, and then use R-NMS to combine them as the final prediction. Finally, the final results are obtained by submitting the final predictions to the official DOTA evaluation server.

For HRSC2016, ResNet-101 is selected as the backbone. The short side of the image in HRSC is resized to 800 and the long side is adaptively resized according to the aspect ratio of the original image for both training and testing datasets. We train the model for total 24 epochs with the same batch size. The initial learning rate is 0.005 and is decreased by 0.1 at epoch 16 and 22. The instances in HRSC have large aspect ratio, so we change the base anchor scales to [32] and change the anchor ratios to [1/3, 1/2, 1, 2, 3] to cover the size of all objects as much as possible. For data augmentation, each image is rotated 90 degrees, 180 degrees, 270 degrees in turn before training.

For both DOTA and HRSC dataset, we use the stochastic gradient descent (SGD) with momentum for network optimization, the weight decay is 0.0001 and momentum is 0.9.

4.4. Ablation studies

In this subsection, we perform several ablation studies to analyze and discuss the impact of the proposed Cascading Attention Network (CA-Net), Adaptive Feature Concatenate Network (AFC-Net) and OBB Multi-Definition and Selection Strategy (OBB-MDS-Strategy) over the DOTA split validation dataset.

Baseline setup. The Faster R-CNN [14] is extended to detect oriented objects, which is used as the baseline for ablation experiments. Jiang et al. [2] prove that adding an extra HBB branch can

improve the OBB detection performance, so we add an extra HBB branch to form the baseline. ResNet-50-FPN is used as the backbone. As Table 1 shows, the baseline gets 68.03% mAP for oriented bounding box task in our implementation.

The impact of OBB-MDS-Strategy. Regression-based object detection methods have similar feature error problem because of the specificity of the angle definition in OpenCV five-parameter system. We design the OBB-MDS-Strategy to solve this problem. According to Table 1, we can observe that the detection performance of mAP is increased by 0.91% on the OBB task. The detection performance of objects with arbitrary directions has been significantly improved, such as bridge, ship, soccer-ball field, harbor and helicopter. This demonstrates that by making the representation of the five parameters of GT OBB consistent with the representation of five parameters of the predicted OBB, the novel OBB selection strategy effectively eliminates similar feature error. Fig. 8. shows some wrong prediction OBBs caused by similar feature error and the correct OBBs revised by the OBB-MDS-Strategy.

We also compare our OBB-MDS-Strategy with the similar method introduced in [5]. Ref. [5] proposes an IoU-smooth L1 loss, which suppresses the process of OBB regression by multiplying Smooth L1 loss and an IoU factor. Table 2 compares our proposed method with IoU-smooth L1 loss, and the results show that our method has a better performance.

The ablation experiments above prove that OBB-MDS-Strategy is effective for improving the performance of rotated object detector, without of adding complex structures and number of parameters.

The impact of CA-Net. As discussed in Section 3.1, the CA-Net is designed to guide the network focus on the object areas and suppress the effect of background noises. Table 1 shows that compared with baseline+OBB-MDS-Strategy, the whole CA-Net yields 0.84% improvement on the OBB task. It has a significant effect on improving the detection performance of objects surrounded by complex background such as small vehicle, large vehicle, baseball diamond, basketball court, storage tank, roundabout and harbor.

CA-Net consists of the Patching Self-Attention Module (PSA-Module) and the Supervised Position Attention Module (SPA-Module), so we construct experiments to show the impact of these two modules respectively. To simplify the experiment, we remove the HBB branch as well as other components in the baseline. The performance of each attention module is shown in Table 3. Table 3 shows that single PSA-Module can achieve an improvement by 1.49%, and single SPA-Module can achieve an improvement by 1.01%. After combining them together, mAP increases by 2.05%. We analyze this because the PSA-Module can learn the relationship

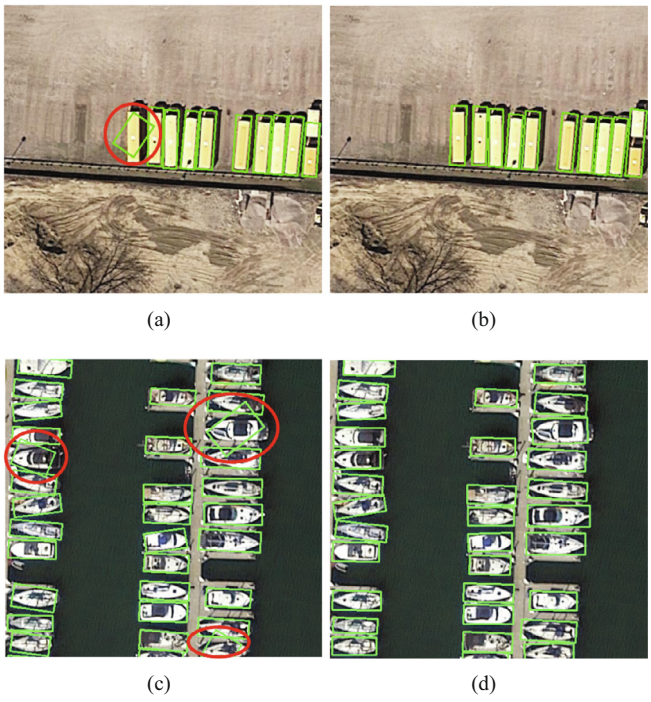

Fig. 8. The visual impact of OBB-MDS-Strategy. (a) A wrong prediction OBB of a large vehicle inside the red ellipse. (b) The correct OBB of large vehicle after introducing the OBB-MDS-Strategy and retraining the network. (c) Wrong prediction OBBs of ships inside the red ellipses. (d) The correct OBBs of ships after introducing the OBB-MDS-Strategy and retraining the network.

Table 2
Comparison of mAP between OBB-MDS-Strategy and IoU-smooth L1 loss.

Method	mAP@OBB(%)
baseline	68.03
+IoU-smooth L1 loss[5]	68.47
+OBB-MDS-Strategy	**68.94**

Table 1
Comparative experiment of our method with ResNet-50 for oriented object detection on DOTA split validation dataset. The images of validation dataset are split as 1024 × 1024 size.

Baseline	Plane	BD	Bridge	GTF	SV	LV	Ship	TC
	89.19	71.48	40.93	65.10	63.64	76.13	78.63	90.80
	BC	ST	SBF	RA	Harbor	SP	HC	mAP
	54.15	80.42	67.18	67.46	64.56	62.48	48.24	68.03
+AFC-Net	Plane	BD	Bridge	GTF	SV	LV	Ship	TC
	89.12	73.21	40.93	67.52	64.33	75.77	78.13	90.74
	BC	ST	SBF	RA	Harbor	SP	HC	mAP
	53.40	80.48	66.80	68.18	64.88	64.05	54.40	68.80
+OBB-MDS-Strategy	Plane	BD	Bridge	GTF	SV	LV	Ship	TC
	89.10	72.13	41.20	65.71	63.79	75.71	**85.83**	90.64
	BC	ST	SBF	RA	Harbor	SP	HC	mAP
	52.46	80.24	72.50	67.47	65.43	60.63	51.29	68.94
+OBB-MDS-Strategy+CA-Net	Plane	BD	Bridge	GTF	SV	LV	Ship	TC
	88.86	74.49	40.49	66.20	**64.32**	76.82	78.69	90.74
	BC	ST	SBF	RA	Harbor	SP	HC	mAP
	55.90	**86.10**	69.02	**69.86**	**71.48**	64.75	49.09	69.78
+OBB-MDS-Strategy+CA-Net+AFC-Net	Plane	BD	Bridge	GTF	SV	LV	Ship	TC
	88.89	**74.58**	**42.23**	**67.78**	64.25	**77.48**	79.15	**90.82**
	BC	ST	SBF	RA	Harbor	SP	HC	mAP
	54.58	85.91	**75.68**	68.90	65.26	**64.76**	**57.06**	**70.49**

Table 3
Ablation study of each attention module in CA-Net. $*$ represents the baseline without HBB branch. SPA-Module† means SPA-Module without supervised learning.

Baseline*	PSA-Module	SPA-Module	SPA-Module†	mAP@OBB (%)
✓	–	–	–	66.63
✓	✓			68.12
✓		✓		67.64
✓	✓		✓	67.55
✓	✓	✓		**68.68**

between objects and their nearby surroundings, while SPA-Module has further suppressed the background parts that PSA-Module misses. We also conduct an experiment to verify the effectiveness of the introduced mask information. Table 3 shows the supervised learning of the mask increases the performance of CA-Net by 1.13%. Fig. 9 shows the comparison between the inputs of CA-Net $\{P2 - P6\}$ and the outputs of CA-Net $\{E2 - E6\}$, indicating that the CA-Net effectively suppress the influence of the background.

We also compare our PSA-Module with the original self-attention module in terms of memory, running time and mAP, as can be seen in Table 4. "SA-Module" means that we use the original self-attention module to replace the PSA-Module unit in the Cascading Attention Network. Running time refers to the time for the attention structure to refine the feature map of the same size 64×64. The results are obtained by inserting these two models into the overall model respectively. Table 4 shows that our proposed PSA-Module spends less memory and time than the original self-attention module, and has a higher mAP, e.g., 5981 versus 6173 for the memory consumption, 0.52 versus 1.27 for the running time, 70.49 versus 69.90 for the mAP. Our PSA-Module splits the feature map into four patches to establish rich relationships between objects and its nearby background. While the original self-attention module establish context relationships on global features, which consumes more time and considers many irrelevant background information far away from objects. The number of patches in PSA-Module also affects the performance. Table 4 shows that the large number of patches can bring fast speed and less memory consumption, but will reduce the mAP. This is because more divided patches means more pixels on the patch boundary; these pixels cannot capture the relationship between each other, which damages the feature representations.

In CA-Net, similar to the Balanced Feature Pyramid module in Libra-rcnn [35], we use a balanced semantic feature which integrates multi-level features to improve the detection performance. In order to explore the best fusion manner, we construct the experiment to integrate the feature pyramid into different levels. We do not consider $P2$ because it will cause memory overflow when sending to PSA-Module. As the Table 5 shows, integrating into $P4$ can achieve the best performance of 68.68%, which is 2.05% larger than the baseline model. The main reason why the way of integration into $P4$ has the best performance is that as the middle level of five feature pyramids in FPN, $P4$ can balance the fine details in low-level features and the semantic information in high-level features. In the Balanced Feature Pyramid in Libra-rcnn [35], the non-local module is used to capture long-range dependencies for each pixel in a balanced semantic feature map. Considering that the correlation between two pixels that are too far apart in a huge aerial image is weak, our PSA-Module splits the balanced semantic feature into four patches to capture local information. In addition, our SA-Module in CA-Net can further suppress the background noises to reduce false alarms. Table 5 shows that our CA-Net performs better than Balanced Feature Pyramid, e.g., 68.68% versus 67.36%. This demonstrates that the correlation between two pixels that are too far apart in a huge aerial image is not strong, and PSA-Module can capture local information by splitting the balanced semantic feature into four patches. In addition, SA-Module in CA-Net can further suppress the background noises to reduce false alarms.

The impact of AFC-Net. AFC-Net is designed to better handle huge scale changes in aerial images. Through the experimental results in Table 1, we can observe that AFC-Net further increases mAP by 0.71% on OBB task. Performance improvements are mainly focused on various-scale objects such as baseball diamond, bridge, ground track field, soccer-ball field, swimming pool and helicopter.

We also perform ablation experiments on AFC-Net alone and prove the effectiveness of AFC-Net in detecting multi-scale objects. According to the description of DOTA dataset in [10], all instances can be divided into three groups according to their height of hori-

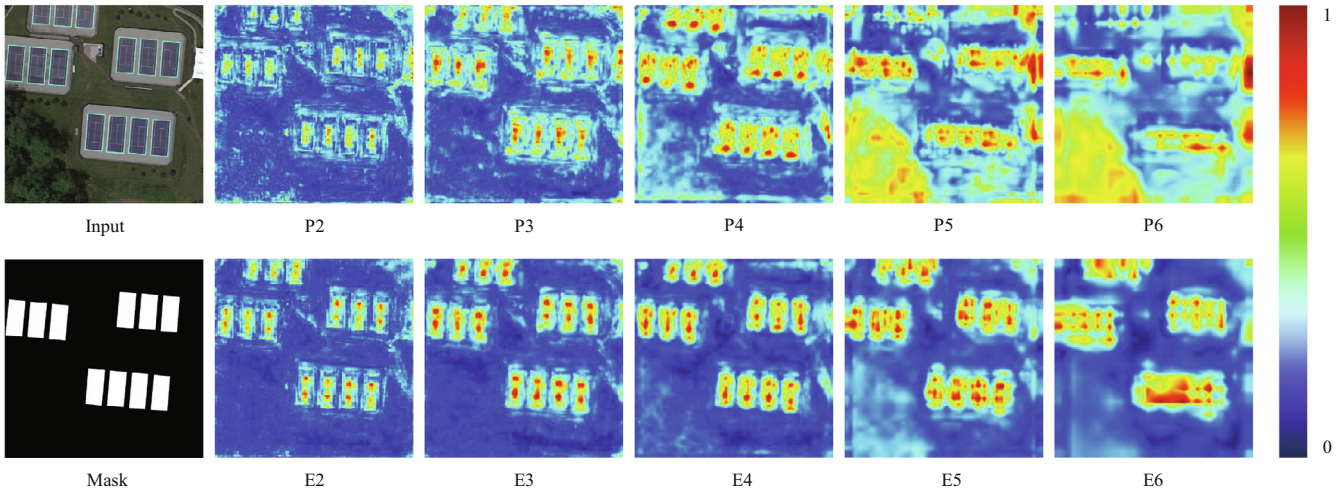

Fig. 9. Visualization of features. The first row feature maps $\{P2 - P5\}$ are the original outputs of FPN. The second row feature maps $\{E2 - E5\}$ are the outputs of CA-Net. As shown in these figures, the background noise is distinctly suppressed in the output feature map of CA-Net, especially at high level E5.

Table 4
Computational requirement comparisons with different setting of the Cascading Attention Network on DOTA split validation set. The number in brackets indicates the number of patches in PSA-Module.

Method	Size of P4	Memory (MB)	Running Time (ms)	mAP@OBB (%)
SA-Module [38] + SPA-Module	64 × 64	6173	1.27	69.90
PSA-Module (4) + SPA-Module	64 × 64	5981	0.52	**70.49**
PSA-Module (16) + SPA-Module	64 × 64	5935	0.29	70.27
PSA-Module (64) + SPA-Module	64 × 64	**5922**	**0.28**	69.51

Table 5
Different fusion manners of the multi-scale feature maps for Cascading attention network (CA-Net) and the comparison between CA-Net and Balanced Feature Pyramid.

Method	mAP@OBB (%)
Baseline*	66.63
+Integration into P3	68.17
+Integration into P4	**68.68**
+Integration into P5	68.19
+Integration into P6	67.81
+Balanced Feature Pyramid [35]	67.36

Table 6
Performance of AFC-Net. Subscripts S and L refer to small and large instances in DOTA validation set, respectively.

Method	mAP@OBB (%)	AP_S (%)	AP_L (%)
Baseline (FPN-based)	68.03	42.75	63.89
+AFC-Net	**68.80**	**43.04**	**64.94**

zontal bounding box: small for range [10,50), middle for range [50,300), and large for range above 300. The number of large objects accounts for only 4% of the total amount of instances in the validation dataset, so we consider the middle group as the large group, too. From the results in Table 6, it can be seen that the detection performance of mAP is increased by 0.77% on the OBB task, the mAP of small group and large group is increased by 0.29% and 1.05%, respectively.

These experimental results demonstrate that aggregating information of different scales according to their importance is effective to improve the detection accuracy of multi-scale objects.

4.5. Comparison with other methods

In this subsection, we compare the performance of our proposed method with the representative methods on the DOTA dataset and HRSC2016 dataset. We obtain the results of DOTA by submitting our predictions to the official DOTA evaluation server.

4.5.1. Results on DOTA

We compare our method with other different methods on OBB and HBB tasks of DOTA dataset in Tables 7 and 8. Here, we use the pre-trained models Resnet-50 and Resnet-101 as our backbone. When the backbone is Resnet-101, due to GPU memory limitations, we have reduced the number of images per GPU.

Results on OBB Task. As can be seen in Table 7, we compare with SSD [8], YOLOv2 [17], FR-O [10], R-DFPN [23], R^2CNN, RRPN [3], ICN [25], RADet [6], RoI-Transformer [24], SCRDet [5] and SegmDet [26]. The best result is highlighted in bold. Among these methods, both ICN and RoI-Transformer generate multi-oriented proposals to better fit the arbitrary-oriented objects region, achieving 68.20% and 69.56% mAP respectively. Both RADet and SegmRDet gain the oriented bounding box of the objects through the mask predicted by the network, achieving 69.09% and 74.14% mAP respectively. By using a novel image synthesizing based training data augmentation technology, SegmRDet has the best performance in large-vehicle (LV), roundabout (RA), harbor and swimming pool (SP). SCRDet and our proposed AMFFA-Net output oriented bounding boxes by regressing coarse horizontal region proposals, which gain performance at 72.61% and 76.27%. For plane, Tennis court (TC), basketball court (BC),storage tank (ST) and Soccer-ball field (SBF), SCRDet has the best performance. Our AMFFA-Net achieves the best performance in baseball diamond (BD), bridge, ground track field (GTF), small vehicle (SV) (78.71% with Resnet-50), ship, and helicopter (HC). Just for the two categories of large-vehicle (LV) and roundabout (RA), the APs of our model have 3.98% and 3.77% lower than that of the best results. For the other categories, our method has achieved better or similar APs. These results are all within reasonable ranges. Compared with SegmRDet, our AMFFA-Net focuses on structural innovation. Firstly, the CA-Net composed of two attention modules can largely suppress the background noises. Then, the AFC-Net can adaptively stack the feature maps for dealing with the multi-scale change of objects. Lastly, the OBB-MDS-Strategy can regress rotated bounding boxes more smoothly. Our method can achieve the highest 76.27% mAP on the OBB task of DOTA dataset under different measures. Some qualitative results of AMFFA-Net on DOTA are given in Fig. 10.

Table 7
Comparative experiment of our method for oriented object detection on DOTA.

method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP
SSD [8]	57.85	32.79	16.14	18.67	0.05	36.93	24.74	81.16	25.10	47.47	11.22	31.53	14.12	9.09	0.00	29.86
YOLOv2 [17]	76.90	33.87	22.73	34.88	38.73	32.02	52.37	61.65	48.54	33.91	29.27	36.83	36.44	38.26	11.61	39.20
FR-O [10]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.45	2.52	46.69	44.80	46.30	52.93
R-DFPN [23]	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R^2CNN [2]	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN [3]	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN [25]	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RADet [6]	79.45	76.99	48.05	65.83	65.46	74.40	68.86	89.70	78.14	74.97	49.92	64.63	66.14	71.58	62.16	69.09
RoI-Trans. [24]	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [5]	**89.98**	80.65	52.09	68.36	68.36	60.32	72.41	**90.85**	**87.94**	**86.86**	**65.02**	66.68	66.25	68.24	65.21	72.61
SegmRDet [26]	89.91	80.31	52.85	70.72	76.72	**79.43**	86.33	90.65	82.91	87.1	48.41	**69.29**	**74.68**	**72.36**	50.48	74.14
Ours (ResNet-50)	89.78	81.68	51.03	76.14	**78.71**	70.11	85.08	90.82	84.74	85.77	58.02	67.49	70.70	70.26	63.80	74.94
Ours (ResNet-101)	89.65	**84.20**	**53.06**	**77.62**	74.66	75.45	**87.08**	90.84	86.84	85.33	63.76	65.52	72.26	70.64	**67.13**	**76.27**

Table 8
Comparative experiment of our method with ResNet-101 for horizontal object detection on DOTA.

method	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP
SSD [8]	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01	10.59
YOLOv2 [17]	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.35	34.65	16.02	37.62	47.23	25.5	7.45	21.39
R-FCN [15]	81.01	58.96	31.64	58.97	49.77	45.04	49.29	68.99	52.07	67.42	41.83	51.44	45.15	53.30	33.89	52.58
FR-H [10]	80.32	77.55	32.86	68.13	53.66	52.49	50.04	90.41	75.05	59.59	57.00	49.81	61.69	56.46	41.85	60.46
FPN [9]	88.70	75.10	52.60	59.20	69.40	78.80	84.50	90.60	81.30	82.60	52.50	62.10	6.60	66.30	60.10	72.00
ICN [25]	90.00	77.70	53.40	73.30	73.50	65.00	78.20	90.80	79.10	84.80	57.20	62.10	73.50	70.20	58.10	72.50
IoU-Adaptive [48]	88.62	80.22	53.18	66.97	**76.30**	72.59	84.07	90.66	80.95	76.24	57.12	**66.65**	74.08	66.36	56.85	72.72
SCRDet [5]	**90.18**	81.88	55.30	73.29	72.09	77.65	78.06	**90.91**	82.44	**86.39**	64.53	63.45	75.77	**78.21**	60.11	75.35
Ours (ResNet-101)	89.67	**84.95**	**59.50**	**76.96**	75.53	**79.86**	**87.61**	90.86	**87.25**	85.66	**65.10**	66.07	**79.35**	77.75	**64.76**	**78.06**

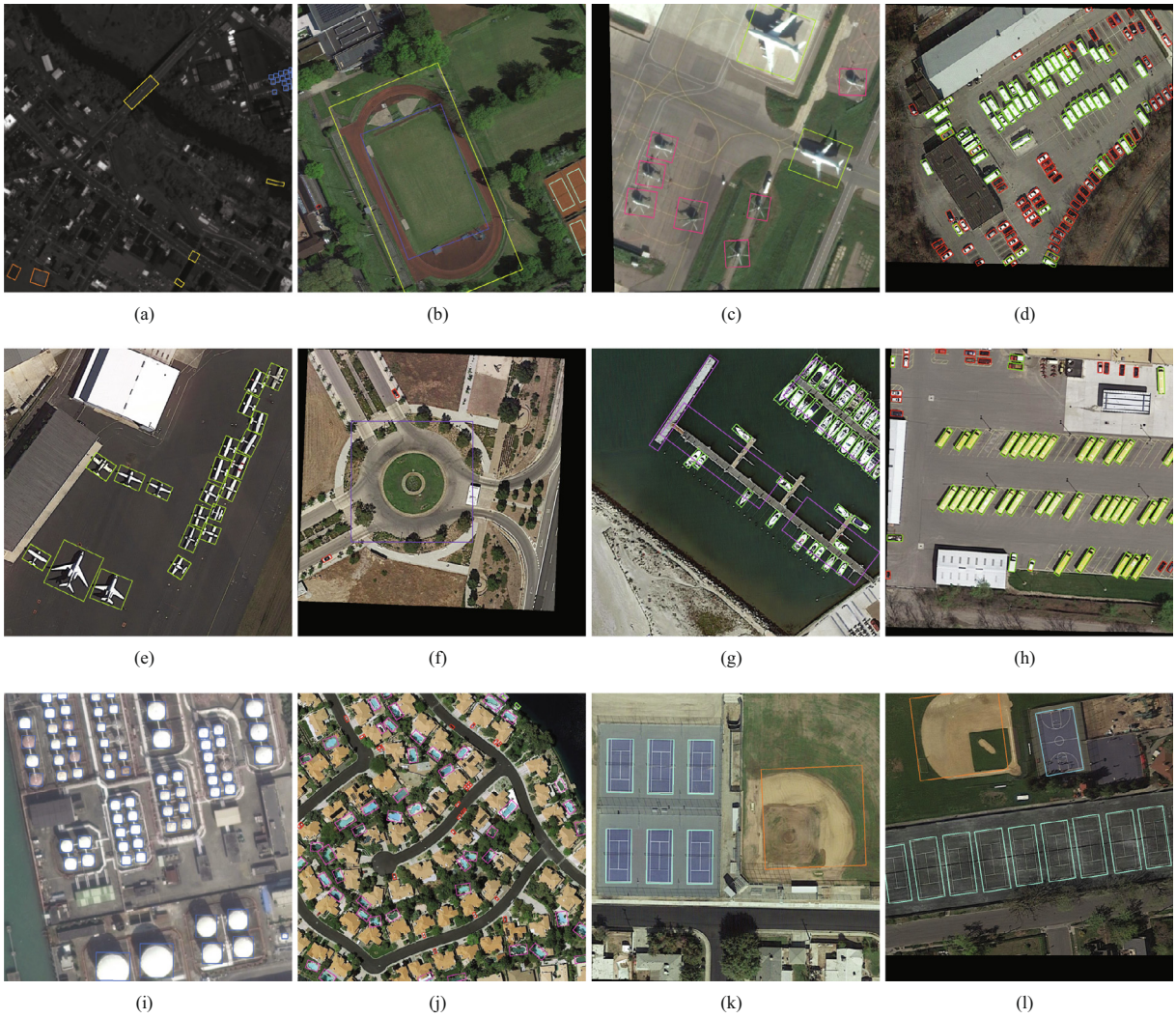

Fig. 10. Some detection results of our method on the DOTA dataset. (a) Detection results of Bridge, Baseball diamond (BD) and Storage tank (ST). (b) Detection results of Ground track field (GTF) and Soccer-ball field (SBF). (c) Detection results of Helicopter (HC) and Plane. (d) Detection results of Large vehicle (LV) and Small vehicle (SV). (e) Detection results of Plane (PL). (f) Detection results of Roundabout (RA) and Small vehicle (SV). (g) Detection results of Ship and Harbor. (h) Detection results of Small vehicle (SV) and Large vehicle (LV). (i) Detection results of Storage tank (ST). (j) Detection results of Swimming pool (SP). (k) Detection results of Tennis court (TC) and Baseball diamond (BD). (l) Detection results of Tennis court (TC), Baseball diamond (BD) and Basketball court (BC).

Results on HBB Task. In order to better verify the performance of our method, we also conduct experiment on the HBB task of DOTA dataset. For HBB task, we compare our method with SSD [8], YOLOv2 [17], R-FCN [15], FR-H [10], FPN [9], ICN [25], IoU-Adaptive [48], and SCRDet [5]. From the results shown in Table 8, we can see that compared with other methods, our method has a relatively better performance in the detection of baseball-diamond (BD), bridge, ground track field (GTF), large vehicle (LV), ship, basketball court (BC), soccer-ball-field (SBF), harbor and helicopter. The mAP of our method on the HBB task is 78.06%, outperforms other compared methods. The result show that our method can also achieve high performance in HBB detection.

Table 9
Results on HRSC2016.

Method	Backbone	mAP (%)	Speed
R^2CNN [2]	ResNet101	73.07	2fps
RRPN [51]	ResNet101	79.08	3.5fps
SCRDet [5]	-	83.41	5fps
RRD [20]	VGG16	84.3	slow
RoI-Transformer [24]	ResNet101	86.20	6fps
R^3Det [49]	MobileNetV2	86.67	**20fps**
MFIAR-Net [50]	ResNet-101	89.81	7.1fps
Ours	ResNet-101	**89.90**	12.6fps

4.5.2. Results on HRSC2016

HRSC dataset contains a lot of ship instances with large aspect ratio and arbitrary orientation, which poses a great challenge to rotation detectors. Table 9 shows the quantitative comparisons between our method and other methods on the HRSC2016 dataset. In order to make a fair comparison, we set the image size with 800×800 and test the single image with post process (i.e. R-NMS) when calculating speed. Compared with other methods, we have achieved competitive performance. R^3Det [49] designs a feature refinement module to solve feature misalignment problem in existing refined single-stage detector, and achieve 86.67% accuracy and 20 fps speed when the input image size is 600×600 and the backbone is MobileNetv2; MFIAR-Net [50] uses a double-path feature attention network to guide the network to focus on object regions and achieve 89.81% accuracy close to ours. But our method is 5 fps faster than MFIAR-Net. Compared with other methods, our method has a balance of speed and accuracy, reaching the highest 89.90% mAP and the second fast 12.6 fps. This result proves that our method can achieve both high mAP and fast speed on HRSC2016.

5. Conclusion

In this paper, we have presented an *Adaptive Multi-Level Feature Fusion and Attention-Based Network* (AMFFA-Net), which is designed for detecting objects commonly seen in remote sensing images with huge orientation variations, complicated and large background, and wide multi-scale distribution. Firstly, a Cascading Attention Network combining two attention modules is proposed to reduce the effects of background noises, making the network focus more on the region of interest. Secondly, we present an Adaptive Feature Concatenate Network that can adaptively integrate features pooled from multi-size feature levels to handle the wide multi-scale distribution problem in aerial images. Thirdly, an OBB Multi-Definition and Selection Strategy is proposed to change the definition of the original GT OBB in the training process for the purpose of dealing with the huge orientation variations problem. Experimental results on public datasets including DOTA and HRSC2016 show that our framework outperforms the compared methods. Based on the good performance, we argue that the proposed methods have high potential to address other detection problems. In the future, we aim to reduce the parameters of our network to speed up the training process, and use more advanced backbone, baseline and richer data augmentation methods to further improve the detection performance.

CRediT authorship contribution statement

Luchang Chen: Conceptualization, Methodology, Software, Investigation, Visualization, Writing - original draft, Writing - original draft. **Chunsheng Liu:** Conceptualization, Methodology, Validation, Formal analysis, Writing - review & editing. **Faliang Chang:** Supervision, Writing - review & editing. **Shuang Li:** Writing - review & editing. **Zhaoying Nie:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key R&D Program of China (NO. 2018YFB1305300), Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO. 2019JZZY010130, 2020CXGC010207), and China Natural Science Foundation Committee (61673244, 61703240).

References

[1] K. Li, G. Wan, G. Cheng, L. Meng, J. Han, Object detection in optical remote sensing images: A survey and a new benchmark, ISPRS Journal of Photogrammetry and Remote Sensing 159 (2020) 296–307.
[2] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, Z. Luo, R2CNN: Rotational region CNN for orientation robust scene text detection, arXiv:1706.09579 (2017)..
[3] Z. Liu, J. Hu, L. Weng, Y. Yang, Rotated region based cnn for ship detection, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, 2017, pp. 900–904..
[4] G. Zhang, S. Lu, W. Zhang, Cad-net: A context-aware detection network for objects in remote sensing imagery, IEEE Transactions on Geoscience and Remote Sensing 57 (12) (2019) 10015–10024..
[5] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, K. Fu, Scrdet; Towards more robust detection for small, cluttered and rotated objects, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8232–8241.
[6] Y. Li, Q. Huang, X. Pei, L. Jiao, R. Shang, Radet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images, Remote Sensing 12 (3) (2020) 389..
[7] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, J. Yang, Feature-attentioned object detection in remote sensing imagery, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3886–3890.
[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[10] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Dota: A large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3974–3983.
[11] Z. Liu, H. Wang, L. Weng, Y. Yang, Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds, IEEE Geoscience and Remote Sensing Letters 13 (8) (2016) 1074–1078.
[12] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[13] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[14] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[15] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems, 2016, pp. 379–387.
[16] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.
[17] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[19] M. Liao, B. Shi, X. Bai, Textboxes++: A single-shot oriented scene text detector, IEEE transactions on image processing 27 (8) (2018) 3676–3690..
[20] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, X. Bai, Rotation-sensitive regression for oriented scene text detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5909–5918.
[21] X. Zhao, C. Zhao, H. Guo, Y. Zhu, M. Tang, J. Wang, Elite loss for scene text detection, Neurocomputing 333 (2019) (2019) 284–291.

[22] L. Liu, Z. Pan, B. Lei, Learning a rotation invariant detector with rotatable bounding box, arXiv:1711.09405 (2017)..
[23] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, Z. Guo, Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks, Remote Sensing 10 (1) (2018) 132..
[24] J. Ding, N. Xue, Y. Long, G. S. Xia, Q. Lu, Learning roi transformer for oriented object detection in aerial images, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2019-June, 2019, pp. 2844–2853..
[25] S.M. Azimi, E. Vig, R. Bahmanyar, M. Körner, P. Reinartz, Towards multi-class object detection in unconstrained remote sensing imagery, in: Asian Conference on Computer Vision, Springer, 2018, pp. 150–165.
[26] Y. Wang, L. Wang, H. Lu, Y. He, Segmentation based rotated bounding boxes prediction and image synthesizing for object detection of high resolution aerial images, Neurocomputing 388 (2020) 202–211.
[27] K. Fu, Z. Chang, Y. Zhang, X. Sun, Point-based estimator for arbitrary-oriented object detection in aerial images, IEEE Transactions on Geoscience and Remote Sensing (2020) 1–18.
[28] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 734–750.
[29] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, arXiv:1904.07850 (2019)..
[30] X. Zhou, J. Zhuo, P. Krahenbuhl, Bottom-up object detection by grouping extreme and center points, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 850–859.
[31] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9627–9636.
[32] Z. Xiao, L. Qian, W. Shao, X. Tan, K. Wang, Axis learning for orientated objects detection in aerial images, Remote Sensing 12 (6) (2020) 908..
[33] Y. Lin, P. Feng, J. Guan, Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection, arXiv:1912.00969 (2019)..
[34] H. Chen, L. Zhang, J. Ma, J. Zhang, Target heat-map network: An end-to-end deep network for target detection in remote sensing images, Neurocomputing 331 (2019) 375–387.
[35] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra r-cnn: Towards balanced learning for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 821–830.
[36] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems 28, 2015, pp. 2017–2025.
[37] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[38] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 7354–7363.
[39] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[40] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[41] J. Wang, Y. Yuan, G. Yu, Face attention network: An effective face detector for the occluded faces, arXiv:1711.07246 (2017)..
[42] A. Li, J. Qi, H. Lu, Multi-attention guided feature fusion network for salient object detection, Neurocomputing 411 (2020) 416–427.
[43] C. Liu, F. Chang, Hybrid cascade structure for license plate detection in large visual surveillance scenes, IEEE Transactions on Intelligent Transportation Systems 20 (6) (2018) 2122–2135.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[45] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, Pdnet: Prior-model guided depth-enhanced network for salient object detection, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 199–204.
[46] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8026–8037.
[48] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, H. Li, Iou-adaptive deformable r-cnn: Make full use of iou for multi-class object detection in remote sensing imagery, Remote Sensing 11 (3) (2019) 286..
[49] X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, G. Yu, R3det: Refined single-stage detector with feature refinement for rotating object, arXiv:1908.05612 (2019)..
[50] F. Yang, W. Li, H. Hu, W. Li, P. Wang, Multi-scale feature integrated attention-based rotation network for object detection in vhr aerial images, Sensors 20 (6) (2020) 1686..
[51] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Transactions on Multimedia 20 (11) (2018) 3111–3122.

Luchang Chen received the B.S. degree in automation from Jilin University, Changchun, China, in 2019. He is currently pursuing the M.S. degree in the school of control science and engineering at Shandong University, Jinan, China. His research interests include computer vision, deep learning, and object detection.

Chunsheng Liu received M.S. degree and Ph.D. degree in Pattern Recognition and Machine Intelligence from Shandong University, Jinan, China in 2012 and 2016, respectively. He was a postdoctor and visiting researcher in University of Washington from 2018 to 2019. He is currently an associate professor at School of Control Science and Engineering, Shandong University. His research interests include pattern recognition, machine learning, intelligent transportation system, and bionic intelligence.

Faliang Chang received his B.S. and M.S. degrees from Shandong Polytechnic University, Jinan, China in 1986 and 1989, respectively. He received his Ph.D. degree in Pattern recognition and Intelligence Systems, Shandong University, in 2003. He has been an professor in Pattern Recognition and Machine Intelligence with the School of Control Science and Engineering, at Shandong University since 2003. Now, his research interests include computer vision, image processing, intelligent transportation system, and multi-camera tracking methodology.

Shuang Li received the B.S. degree in School of Mechanical, Electrical and Information Engineering from Shandong University at Weihai, Weihai, China, in 2010; and the M.S. degree in pattern recognition and intelligent systems from Shandong University, Jinan, China, in 2013. She is pursing the Ph.D. degree in Control Science and Engineering at the School of Control Science and Engineering, Shandong University, Jinan, China. Her current research interests include automatic target detection and recognition, machine learning, deep learning and traffic flow parameter estimation.

Zhaoying Nie received the B.S. degree in automation from Shandong University at Weihai, Weihai, China, in 2019. He is currently an M.S. candidate in the school of control science and engineering at Shandong University, Jinan, China. His research interests include deep learning, reinforcement learning, and scene understanding.