$\mathcal{M}^2$-Net: A Multi-scale Multi-level Feature Enhanced Network for Object Detection in Optical Remote Sensing Images

Xinhai Ye[1],Fengchao Xiong[1], Jianfeng Lu[1], Haifeng Zhao[2], Jun Zhou[3]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2]School of Software Engineering, Jinling Institute of Technology, China
[3]School of Information and Communication Technology, Griffith University, Australia
Emails: {yyxxhh, fcxiong, zhf}@njust.edu.cn, zhf@jit.edu.cn, jun.zhou@griffith.edu.au

***Abstract*—Object detection in remote sensing images is a challenging task due to diversified orientation, complex background, dense distribution and scale variation of objects. In this paper, we tackle this problem by proposing a novel multi-scale multi-level feature enhanced network ($\mathcal{M}^2$-Net) that integrates a Feature Map Enhancement (FME) module and a Feature Fusion Block (FFB) into Rotational RetinaNet. The FME module aims to enhance the weak features by factorizing the convolutional operation into two similar branches instead of one single branch, which helps to broaden receptive field with less parameters. This module is embedded into different layers in the backbone network to capture multi-scale semantics and location information for detection. The FFB module is used to shorten the information propagation path between low-level high-resolution features in shallow layers and high-level semantic features in deep layers, facilitating more effective feature fusion and object detection especially those with small sizes. Experimental results on three benchmark datasets show that our method not only outperforms many one-stage detectors but also achieves competitive accuracy with lower time cost than two-stage detectors.**

***Index Terms*—Convolutional neural network (CNN), object detection, feature fusion, remote sensing image, multi-scale analysis**

I. Introduction

With the fast development of earth observation satellite technology, large amount of high-resolution optical remote sensing images (RSIs) are more easily accessible every day, making it possible to better monitor and understand the earth. Object detection aims at simultaneously determining the location and categories of the object of interests (e.g. plane, vehicle, ship) in the images. It plays an important role in analyzing the RSIs and promoting their usage in real-world applications such as urban planning, traffic management, map production, etc [1].

Recent years have witnessed the considerable achievement on deep convolutional neural networks (CNNs) based object detection in natural images thanks to their superior advantages in feature and image representation [2], [3]. CNN-based frameworks for object detection can be roughly divided into two categories: one-stage methods [4]–[7] and two-stage methods [8]–[11]. One-stage methods such as YOLO [4], YOLOV2 [5], SSD [6] and RetinaNet [7] consider object detection as a regression problem and simultaneously predict object location and object class through an end-to-end structure. Two-stage methods instead divide this task into two steps. First, several regions of interests are produced by a region proposal module, e.g. using select search [8] or region proposal network (RPN) [10]. Then CNN is employed to extract robust features from each region and make class-specific predictions. Two-stage methods usually achieve better accuracy but cost more time on prediction. Representative two-stage detectors include R-CNN [8], Fast R-CNN [9] and Faster R-CNN [10]. Moreover, improvement has also been made on feature pyramid networks (FPN) [12] to support multi-scale detection.

Although object detection methods have achieved very promising performance in natural scene images, it is unrealistic to directly apply these detectors to optical RSIs, which are captured with camera mounted on satellites or aeroplanes [13]. Fig. 1 shows some sample images for remote sensing object detection. Compared with natural object detection, remote sensing object detection has the following unique challenges:

- First, objects in RSIs are usually of small sizes with arbitrary orientations, scale variation and dense distribution, which significantly increase the difficulty of detection.
- Second, remote sensing objects are prone to be overwhelmed by cluttered and complex backgrounds which potentially introduce more false positives and noises.
- Third, RSIs are lack of contrast and texture details, which are very discriminative clues for a detector, leading to limited detection accuracy.

To address the above challenges, many researchers focus on introducing domain-specific knowledge to existing networks [14], [15]. One strategy is to embed rotation-aware

This work was supported in part by the National Natural Science Foundation of China under Grant 62002169, the National Key Research and Development Program of China under Grant 2017YFB1300205, the Research Foundation for Advanced Talents and Incubation Foundation of Jingling Institute of Technology under grant JIT-B-201717 and JIT-FHXM-201808 and the Major Program of University Natural Science Research of Jiangsu Province under grant 16KJA520003. (Corresponding authors: Fengchao Xiong; Jianfeng Lu.)

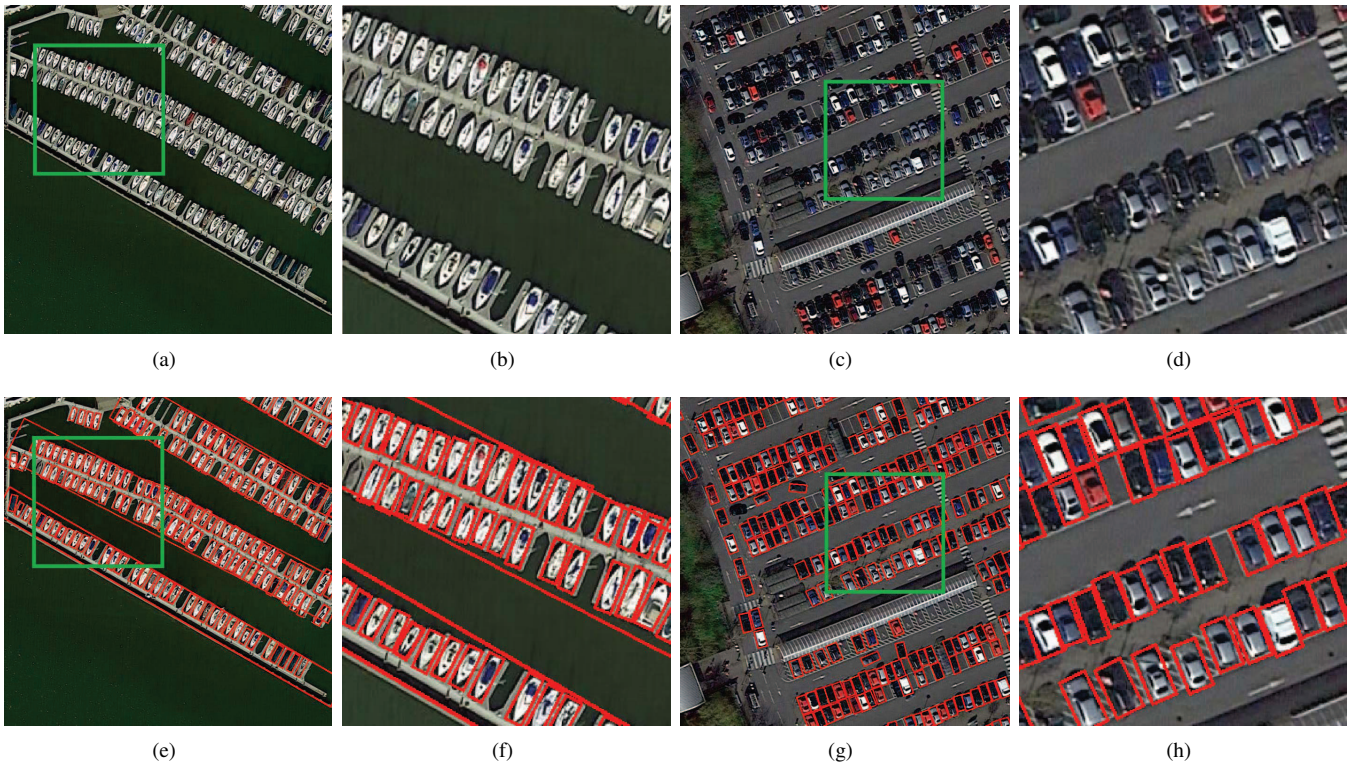


Fig. 1. Sample images in remote sensing object detection. For each scene, the green bounding box on the left image shows a selected region, and the right image shows the magnified view of the region.

prior information into CNN models by introducing additional rotation-invariant layers or rotational region proposal networks [14], [16]–[18]. For example, Cheng *et al.* [18] added a rotation-invariant layer to R-CNN framework to enforce CNN feature representations to share close mapping before and after rotation for object detection. Ding *et al.* developed a lightweight region of interest (RoI) transformer to realize the geometry transformation between horizontal RoIs and rotational ROIs, enabling network to extract rotation-invariant region features for arbitrary-oriented object detection [19].

Furthermore, feature enhancement is also investigated to boost the detection performance, among which attention mechanism [20], [21] and feature fusion [22]–[26] are vastly explored. Attention mechanism is based on the fact that human brain tends to put more concentration on a certain critical region when processing a large amount of perceived information. As for detection in RSIs, attention is helpful to guide the network to focus on prominent regions [27], [28]. Based on Faster R-CNN, multi-scale spatial and channel-wise attention mechanism [29] was proposed to make the detector pay more attention to foreground regions and overcome the influence of the complex background, facilitating precise localization.

Feature fusion exploits the context information for detection by combining the power of low-resolution high-level features from deeper layers with high-resolution low-level features from shallow layers. As a result, the produced features are enriched and enhanced, especially for small objects or occluded objects [15], [28]. Liu *et al.* [24] enhanced YOLOv2 with oriented response dilated convolution and fused feature maps from different layers, enabling to detect objects at multiple scales in complex geospatial images. Driven by the power of FPN in multi-scale detection, a multi-scale rotation dense feature pyramid network was proposed in [30] for ship detection where dense connections were used to enhance propagation and encourage reuse of high-level semantical features from different layers. Alternatively, image cascade network (ICN) [31] combines image cascade and FPN to allow extracting features at different levels and scales. Feature-merged single-shot detection (FMSSD) [15] leverages an atrous spatial feature pyramid (ASFP) to pass the semantic features from a high level to a low level, in which atrous convolutions with multiple rates were adopted to enlarge the receptive field. In [28], Zhang *et al.* proposed a context-aware detection network (CAD-Net) to integrate scene-level global semantics and object-level local contexts of objects for more consideration of low-contrast objects. SCRDet [25] employed a sampling fusion network, which combines feature fusion with effective anchor sampling for improved sensitivity to small objects. Moreover, fast detection based on light-weight backbones is also studied [17], [32].

In this paper, we propose a **M**ulti-scale **M**ulti-level feature enhanced **Net**work ($\mathcal{M}^2$**-Net**), to boost remote sensing object detection. As shown in Fig. 2, it is a one-stage network and inherits from Rotation RetinaNet (RetinaNet-R) [33]. Two additional modules, i.e., Feature Map Enhancement module (FME) and Feature Fusion Block (FFB) are introduced to

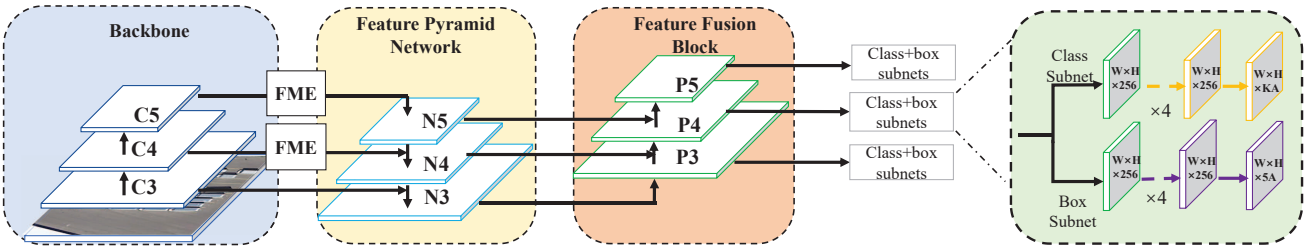


Fig. 2. The architecture of $\mathcal{M}^2$-Net. This network has two additional module named feature map enhancement module (FME) and feature fusion block (FFB). FME is embedded between backbone and feature pyramid network, which aims at enhancing the weak features from the backbone. FFB is set after the feature pyramid network for the purpose of getting more accurate location and semantic information from the backbone.

encourage multi-scale multi-level feature enhancement for more consideration of the unique characteristics of RSIs influencing robust detection. The FME module is embedded into different layers to enhance the weak features from the backbone network, so as to selectively integrate multi-scale features of different semantics and localization information. FFB module aims to simultaneously take better advantages of low-level texture features for accurate localization and high-level semantic features for classification via a bottom-up path augmentation. Experimental results on the DOTA [13], NWPU VHR-10 [18] and UCAS-AOD [34] datasets demonstrate the effectiveness and generalization capability of the proposed method while fast detection speed can be achieved.

The rest of the paper is organized as follows. Section II describes the proposed $\mathcal{M}^2$-Net and analyzes its advantages in remote sensing object detection. Section III presents the experimental results on three widely-used datasets. Section IV concludes the paper with future work.

II. Proposed $\mathcal{M}^2-$Net

In this section, we describe in detail the proposed $\mathcal{M}^2$-Net, including its overall architecture, additive feature map enhancement, feature fusion module and loss function setting.

A. Overview of the Proposed Network

Fig. 2 shows the overall structure of $\mathcal{M}^2$-Net. Inherited from RetinaNet-R, it contains a backbone network, a feature pyramid network as well as a classification and regression subnetwork. In order to achieve rotation invariant detection, five parameters (x, y, w, h, θ) are used to represent arbitrary-oriented rectangle, where x, y, w, h respectively indicates the center coordinates of ground truth box, the width and the height. An angular offset is added to the regression subnet, and the bounding box is defined as follows:

$$\begin{aligned} t_x &= (x - x_a)/w_a,\ t_y = (y - y_a)/h_a \\ t_w &= \log(w/w_a),\ t_h = \log(h/h_a) \\ t_\theta &= \theta - \theta_a \\ t_x^{'} &= (x^{'} - x_a)/w_a,\ t_y^{'} = (y^{'} - y_a)/h_a \\ t_w^{'} &= \log(w^{'}/w_a),\ t_h^{'} = \log(h^{'}/h_a) \\ t_\theta^{'} &= \theta^{'} - \theta_a \end{aligned} \tag{1}$$

where x_a and y_a are the coordinates of the center of the anchor box and $x^{'}$ and $y^{'}$ are the coordinates of the centre of the predicted box, likewise for other parameters.

Due to challenges of objects in RSIs, such as smaller size, scale variation and lower contrast, etc., RetinaNet-R can not be directly applied for remote sensing object detection. To this end, we propose to embed FME and FFB modules into RetinaNet-R, aiming to introduce domain-specific knowledge of geospatial objects for enhanced feature representation so as to improve detection.

B. Feature Map Enhancement (FME)

RSIs suffer from feature size variation and lack of contrast and texture details, which requires the backbone to be equipped with strong feature extraction ability. Generally, the feature extraction ability of backbone can be improved by increasing the width or depth. As we know, backbone such as ResNet [2] uses pooling layers to reduce the resolution, which makes the deep layers get more semantic features and promote the classification. With the increasing of the depth or layers, the spatial resolution of the feature map decreases, hindering capacity of predicting the locations of objects in RSIs, especially for small objects. Therefore, it is not reasonable to add more layers for enhanced feature representation. Alternatively, Inception Network [35] pointed out that the feature extraction capacity can also be strengthened by broadening the network through putting more branches in the same layer.

To this end, inspired by ACNet [36], an FME module is constructed to more effectively capture the location of objects, as shown in Fig. 3. It consists of two similar branches. In the left branch, a 1×1 convolution layer is used to reduce the channels and parameters. Then a 3×3 convolution layer is used to learn more non-linear relations and broaden the receptive field. Meanwhile we factorize the 3×3 convolution operation into a 1×3 layer and a 3×1 layer for keeping receptive field as well as decreasing inference time. The right branch shares the similar architecture as the left branch but reverses the group of 1×3 and 3×1 convolution layers. Additionally, a shortcut connection is adopted to combine the original features and also ease information propagation. In order to capture different scales of semantics and location information, we embed FME to C4 and C5 layers of backbone ResNet, as shown in Fig. 2.

Moreover, instead of ReLU, we choose Gaussian Error Linear Unit (GELU) [37] as an activation function considering

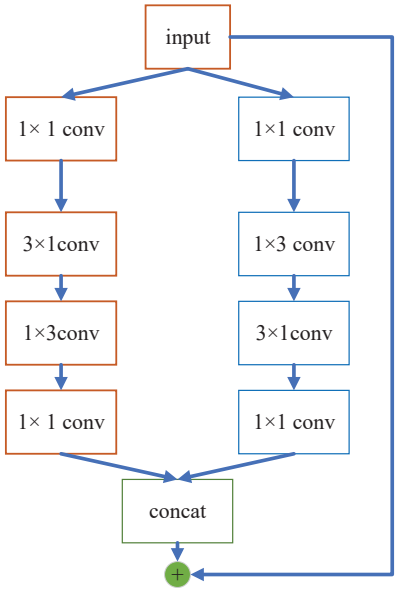


Fig. 3. The architecture of FME.

its powerful capacity of approximate complicated functions and better interpretability. Mathematically, GELU can be approximated with

$$f(x) = 0.5x(1 + \tanh[\sqrt{2/\pi}(x + 0.044715x^3)]) \quad (2)$$

where x is the input.

C. Feature Fusion Block (FFB)

Compared with large image size, e.g. 3000×3000 or more, the sizes of objects in remote sensing image are usually very small, in many cases only covering less than 15 pixels. It is known that the feature maps in deeper layers respond to the high-level semantic signals of entire objects while feature maps in shallow layers are related with low-level localization signals. The long path from shallow layers to deep features weakens and potentially vanishes the accurate localization information of these small targets, significantly reducing detection accuracy. This issue can be overcome by bottom-up path augmentation as in PANet [38] which shortens the information propagation path between deep layers and shallow layers.

Inspired by PANet [38], we introduce feature fusion block (FFB) into detection network, shown in Fig. 4. The P_3 layer of FFB is the same as the N_3 layer of FPN in the original RetinaNet. Each feature map P_i first goes through a 3×3 convolution layer with stride 2, yielding half-size feature map. Then each element of feature map N_{i+1} and the down-sampled map are added through element-wise addition. Subsequently, the P_{i+1} layer is generated by a 3×3 convolution layer after element-wise addition. Same as FME, all convolution layers are followed by a GELU. As shown in Fig. 2, the P_3 layer of FFB comes from the C_3 layer of ResNet where high-resolution information exists. The N_4 and N_5 layers of FPN contain more high-level semantic information. With the FFB module, the low-level localization information is fused with the high-level semantic information for more effective object detection.

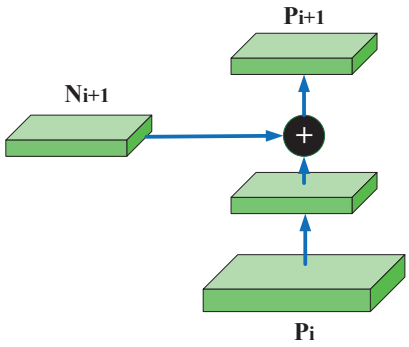


Fig. 4. The architecture of FFB. It comprises of two convolution layers, respectively aiming to reduce the size and adjust the dimension of of feature maps.

D. Loss Function

The same as RetinaNet-R, $\mathcal{M}^2$-Net uses a multi-task focal loss to balance positive and negative samples defined as

$$L = \frac{\lambda_1}{N}\sum_{n=1}^{N} t_n^{'} \sum_{j \in \alpha} L_{reg}(v_{nj}^{'}, v_{nj}) + \frac{\lambda_2}{N}\sum_{n=1}^{N} L_{cls}(p_n, t_n) \quad (3)$$

where $\alpha = (x, y, w, h, \theta)$, N represents the number of anchors. The regression loss L_{reg} is a smooth L_1 loss measuring the differences between ground-truth v_{nj} and predicted one $v_{nj}^{'}$. The focal loss L_{cls} is used for classification. Hyper-parameters λ_1 and λ_2 balance these two losses. Both of them are set to 1 during training.

III. Experiments

In this section, we compare the proposed $\mathcal{M}^2$-Net with several state-of-the-art detectors, including both one-stage and two-stage methods, on both oriented bounding box (OBB) task and horizontal bounding box (HBB) task to demonstrate the advantages of our method. The performance of the competing detectors are extracted from the results reported in the original paper. Since the dataset and experimental setting in those papers and ours are exactly the same, the results are comparable.

A. Experimental Setting

1) Dataset: Three datasets were used for evaluation, including DOTA [13], NWPU VHR-10 [18] and UCAS-AOD [34]. DOTA contains 2806 aerial images with sizes ranging from 800×800 to 4000×4000 pixels. The whole dataset includes 15 categories of objects and 188,282 instances in total. The NWPU VHR-10 contains 800 aerial images, where 650 of them are labeled, covering 10 different categories, all of which are included in DOTA. UCAS-AOD contains 1,510 aerial images with approximate size of 1000×1000. It contains 14,596 instances of planes and cars. Both classes are also included in DOTA. The training set and testing set of DOTA, NWPU VHR-10 and UCAS-AOD are the same as reported in [33], [18] and [13].

TABLE I

RESULT COMPARISON OF OBB TASK ON DOTA DATASET. THE SHORT NAMES ARE DEFINED AS: PL-PLANE, BD-BASEBALL DIAMOND, BR-BRIDGE, GTF-GROUND FIELD TRACK, SV-SMALL VEHICLE, LV-LARGE VEHICLE, SH-SHIP, TC-TENNIS COURT, BC-BASKETBALL COURT, ST-STORAGE TANK, SBF-SOCCER-BALL FIELD, RA-ROUNDABOUT, HA-HARBOR, SP-SWIMMING POOL, HC-HELICOPTER. THE TOP TWO VALUES ARE HIGHLIGHTED IN RED AND BLUE.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP(%)
Two-stage methods																
R-FCN [11]	37.80	38.21	3.64	37.26	6.74	2.60	5.59	22.85	46.93	66.04	33.37	47.15	10.60	25.19	17.96	26.79
FR-O [13]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.4	52.52	46.69	44.80	46.30	52.93
ICN [31]	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
RoI-Transformer [19]	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
CAD-Net [28]	87.80	**82.40**	**49.40**	**73.50**	**71.10**	63.50	76.70	**90.90**	**79.20**	73.30	48.40	60.90	62.00	**67.00**	**62.20**	69.90
SCRDet [25]	**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**
One-stage methods																
SSD [6]	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 [5]	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.25	34.65	16.02	37.62	47.23	25.19	7.45	21.39
Axis-Learning [39]	79.53	77.15	38.59	61.15	67.53	70.49	76.30	89.66	79.07	83.53	47.27	61.01	56.28	66.06	36.05	65.98
RetinaNet-R [33]	88.92	67.67	33.55	56.83	66.11	73.28	75.24	**90.87**	73.95	75.07	43.77	56.72	51.05	55.86	21.46	62.02
$\mathcal{M}^2$-Net	**89.01**	80.02	40.12	68.23	**71.03**	**77.32**	**78.01**	90.82	78.05	77.33	58.02	62.19	**65.55**	61.32	56.32	**70.22**

TABLE II

RESULT COMPARISON OF HBB TASK ON DOTA DATASET. THE TOP TWO VALUES ARE HIGHLIGHTED IN RED AND BLUE.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP(%)
Two-stage methods																
R-FCN [11]	79.33	44.26	36.58	53.53	39.38	34.15	47.29	45.66	47.74	65.84	37.92	44.23	47.23	50.64	34.90	47.24
FR-H [13]	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
ICN [31]	**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
SCRDet [25]	**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**
One-stage methods																
SSD [6]	44.74	11.21	6.22	6.91	2.00	10.24	11.34	15.59	12.56	17.94	14.73	4.55	4.55	0.53	1.01	10.94
YOLOV2 [5]	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
FMSSD [15]	89.11	81.51	48.22	67.94	69.23	73.56	76.87	90.71	**82.67**	73.33	52.65	**67.52**	72.37	**80.57**	**60.15**	72.43
$\mathcal{M}^2$-Net	89.27	**82.63**	**54.02**	72.32	**72.20**	**75.29**	**83.55**	**90.85**	**84.36**	70.85	**59.29**	62.38	**75.07**	71.96	53.79	**73.19**

TABLE III

RESULTS COMPARISON ON NWPU VHR-10 AND UCAS-AOD DATASETS.

Method	Training data	Testing data	mAP(%)
Cheng et al. [18]	NWPU VHR-10	NWPU VHR-10	72.63
ICN [31]	NWPU VHR-10	NWPU VHR-10	95.01
$\mathcal{M}^2$-Net	NWPU VHR-10	NWPU VHR-10	**95.32**
ICN [31]	DOTA	NWPU VHR-10	82.23
$\mathcal{M}^2$-Net	DOTA	NWPU VHR-10	83.12
Xia et al. [13]	UCAS-AOD	UCAS-AOD	89.41
ICN [31]	UCAS-AOD	UCAS-AOD	95.67
$\mathcal{M}^2$-Net	UCAS-AOD	UCAS-AOD	**96.01**
ICN [31]	DOTA	UCAS-AOD	86.13
$\mathcal{M}^2$-Net	DOTA	UCAS-AOD	87.01

TABLE IV

RUNNING TIME OF DIFFERENT METHODS ON DOTA DATASET.

Method	mAP(%)	time(min)
FR-O [13]	52.93	1015
R^{2}CNN [40]	60.67	1043
RoI-Transformer [19]	69.56	1095
RetinaNet-R [33]	59.44	**480**
$\mathcal{M}^2$-Net	**70.22**	503

2) Evaluation Metric: Following the PASCAL VOC 2012 object detection task, we also use mAP to evaluate the detection performance of all methods. Mathematically, mAP is defined by

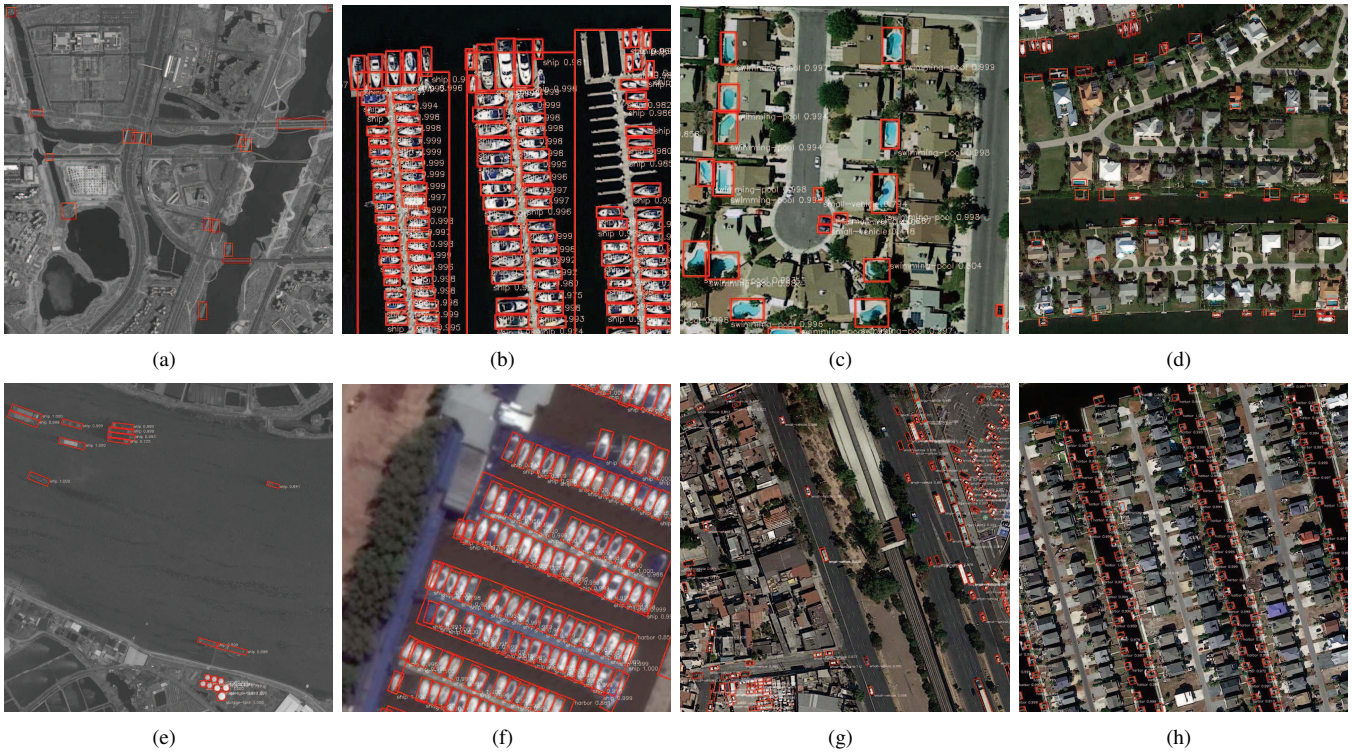
$$\text{mAP} = \frac{1}{C}\sum_{j=1}^{C}\int P_j(R_j)dR_j \tag{4}$$

Here, R_j represents the recall for a given class j of a detector, $P_j(R_j)$ denotes the precision for a given class j when the recall of this class is R_j and C is the number of classes to be detected.

3) Network Settings: ResNet-50 [2] is adopted as the backbone network for feature extraction. We trained the network on a Linux machine with the configuration of one NVIDIA Titan XP GPU and 12GB memory. Stochastic gradient (SGD) with momentum is used for network optimization, whose weight decay and batch size are respectively given by 0.00001 and 1. For DOTA and UCAS-AOD datasets, the learning rate is set to 0.001 and divided by 10 when the number of iterations approaches 360,000 and 480,000 while the total iterations is set to 600,000. In term of the NWPU VHR-10 dataset, the learning rate is set to 0.0001.

B. Comparison of Detection Accuracy

We first report the detection accuracy on DOTA dataset, which includes both OBB task and HBB task. As can be seen in Table I and Table II, methods designed for natural scene images such as FCN, SSD, YOLOV2, provide unsatisfied results due to limited consideration of unique characteristics of RSIs. Among all the compared one-stage methods, the proposed $\mathcal{M}^2$-Net achieves the best detection accuracy by obtaining **70.22**% mAP on OBB task and **73.19**% on HBB task thanks to multi-scale feature learning ability enabled by FME and multi-level feature fusion ability powered by FFB. Compared with the baseline RetinaNet-R, noticeable improvement is made by the proposed $\mathcal{M}^2$-Net. Moreover, our



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Visual results of the proposed $\mathcal{M}^2$-Net on DOTA dataset.(a)-(d): HBB task, (e)-(h): OBB task.

detector wins baseline in almost all the categories. The main reason is that the FME module enhances the weak features and FFB module helps to fuse low-level high-resolutional features and high-level semantic features. Additionally, our detector fails to surpass SCRDet on both tasks. This is because SCRDet is a two-stage detector with RPN which is better at producing rotational anchors, facilitating fitting ground-truth. However, our $\mathcal{M}^2$-Net wins on those objects with small sizes and high density, such as SV, LV and SH. The main reason is that they are usually very small in size and require more information for accurate location and classification, which can be achieved through the proposed FME. In summary, this experiment evidently verifies the effectiveness of proposed $\mathcal{M}^2$-Net in RSIs detection.

Table III shows the detection performance on NWPU VHR-10 and UCAS-AOD datasets, which are respectively used for HBB task and OBB task. In this experiment, we test the detectors on two different training data settings, i.e., DOTA and NWPU VHR-10. As can be seen, our method is also better than the alternatives in both settings, which confirms the superiority and generality of the proposed detector.

C. Visual results

Fig. 5 visualizes the detection results on DOTA dataset. Thanks to the FME module, the visual cues are enhanced, enabling the detector to locate the objects in low contrast scenarios, see Fig. 5(a) and Fig. 5(e). The remaining figures demonstrate the detection performance on objects with dense arrangement, arbitrary rotation and very small size. Thanks to the advantages of FFB in integrating multi-level features at different resolutions, proposed $\mathcal{M}^2$-Net can accurately detect their positions. The superior detection results further verify the effectiveness of proposed method in remote sensing object detection.

D. Comparison of Running Time

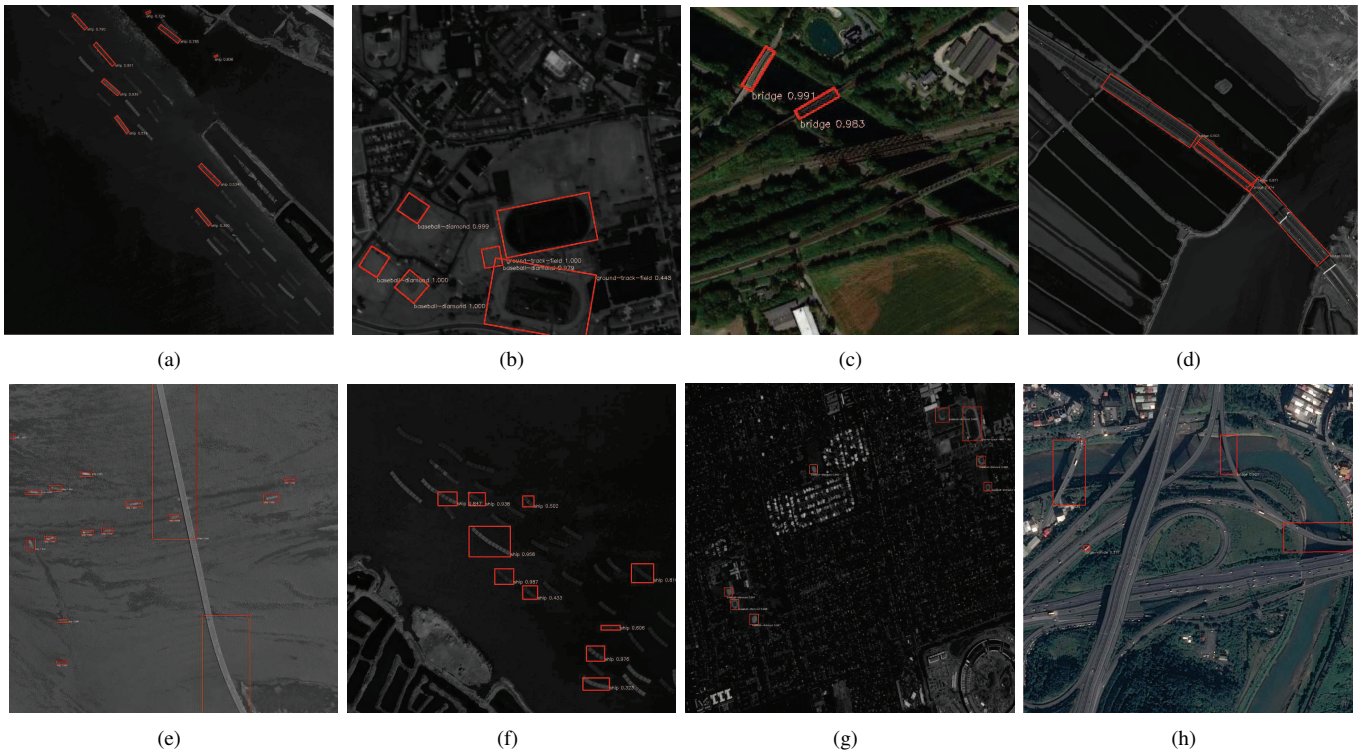
Table IV shows the detection speed on DOTA dataset. We chose FR-O [13], R^{2}CNN [40], RetinaNet-R [33] and RoI-Transformer [19] as the alternative methods for comparison considering their codes are publicly available. We ran all the detectors on a Linux machine with one Titan Xp 12G GPU, two Intel Xeon E5-2620 CPUs and 64G memory. As presented in the table, two-stage detectors use an extra stage RPN to generate proposals, leading to the phenomena that FR-O, R^{2}CNN and ROI-Transformer cost much more time for detection. Compared with the baseline RetinaNet-R, our method cost slightly more time, but obtains a gain of 10.78% mAP. The main reason is that we use a 1×3 convolutional layer and a 3×1 convolutional layer to replace a 3×3 convolutional layer, which can help reduce the inference time. Overall, proposed detector is a time-efficiency detector with competitive detection ability.

E. Ablation study

Table V shows the impact of each proposed component, i.e., FME and FFB, multi-scale (MS) settings and backbone in our pipeline. All experiments were performed on DOTA dataset.

TABLE V
ABLATION STUDY OF COMPONENTS ON DOTA DATASET.

Baseline	Backbone	FME	FFB	MS	mAP (%)@OBB	mAP (%)@HBB
✓	ResNet-50	-	-	-	59.44	62.33
✓	ResNet-50	✓	-	-	61.76	64.02
✓	ResNet-50	-	✓	-	62.92	65.25
✓	ResNet-50	✓	✓	-	64.62	68.21
✓	ResNet-101	✓	✓	-	68.85	71.03
✓	ResNet-101	✓	✓	✓	70.22	73.19



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. Some typical failure predictions produced by our method. (a)-(d): HBB task, (e)-(h): OBB task.

FME. This module aims to enhance the weak features from the backbone to help detect categories with small sizes, high density and low contrast. FME enables to extract multi-scale location information with larger receptive field, yielding a gain of 2.32% and 1.69% respectively on OBB and HBB task. This shows that this module can significantly help the whole network with improved performance.

FFB. This module is a compensate module for the FPN, making the detector get more accurate location information. Thanks to multi-level locational and semantic information powered by FFB, the improved mAP on both OBB and HBB tasks are about 1.2%.

Backbone. Backbone network plays an important role in object detection. Deeper backbone generally indicates more capacity of feature extraction. For this reason, replacing ResNet-50 with ResNet-101 allows our detector gaining 4.23% and 2.82% respectively on OBB and HBB task .

MS. In training step, we resize the images at scales of (0.5,0.6,0.8,1) to increase size diversity. With the help of MS, the mAP reaches 70.22% and 73.19% respectively.

F. Failure cases analysis

Although our method outperforms many detectors, there is room for further exploration. In this section, we analysis the shortcomings of our detector by showing the typical failure cases on DOTA dataset, given in Fig. 6. In RSIs, objects tend to be overwhelmed by complex background, which high likely introduces false positives and noises. As a result, the detector is misled especially in very low-resolution images by missing or falsely detecting the objects. For example, the detector mistakes the embankment for a ship in Fig. 6(a) and Fig. 6(f) and falsely considers the vacant land as a baseball diamond in Fig. 6(b) and Fig. 6(g). The detector fails to detect all the bridges contained in Fig. 6(c) and Fig. 6(h) and only detect parts of the bridge in Fig. 6(d) and Fig. 6(e). This result suggests that our detector should be limited in detecting objects with high aspect ratios which is always a difficult task.

IV. Conclusion

In this paper, we have introduced a novel deep neural network for remote sensing object detection. It contains a feature map enhancement module which enhances weak features in backbone by broadening the network with very few additional parameters, and a feature fusion block which fuses low-level features in shallow network and high-level features, making the detector more powerful of locating small objects. The proposed method was evaluated on DOTA, NWPU VHR-10 and UCAS-AOD datasets. The experimental results show that our method has better capacity of RSIs detection, with higher computational efficiency. In the future, we will extend our method to more discriminative deep models for fine-grained object detection.

References

[1] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.

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

[3] K. He, G. Gkioxari, P. Dollr, and R. Girshick, in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

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

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

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016.

[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[11] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[12] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[13] G. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[14] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, 2019.

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

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

[17] P. Ding, Y. Zhang, W.-J. Deng, P. Jia, and A. Kuijper, "A light and faster regional convolutional neural network for object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 141, pp. 208–218, 2018.

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

[19] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[20] Y. Lin, P. Feng, and J. Guan, "IENet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," *arXiv:1912.00969*, 2019.

[21] C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, "Feature-attentioned object detection in remote sensing imagery," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019.

[22] X. Zhang, K. Zhu, G. Chen, X. Tan, L. Zhang, F. Dai, P. Liao, and Y. Gong, "Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network," *Remote Sensing*, vol. 11, no. 7, p. 755, 2019.

[23] H. Qin, Y. Li, J. Lei, W. Xie, and Z. Wang, "A specially optimized one-stage network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, 2020 (In Press).

[24] W. Liu, L. Ma, J. Wang, and H. Chen, "Detection of multiclass objects in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 791–795, 2019.

[25] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[26] H. Tayara and K. T. Chong, "Object detection in very high-resolution aerial images using one-stage densely connected feature pyramid network," *Sensors*, vol. 18, no. 10, p. 3341, 2018.

[27] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 310–314, 2019.

[28] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10 015–10 024, 2019.

[29] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, 2020.

[30] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and 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*, vol. 10, no. 1, 2018.

[31] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. Asian Conf. on Comput. Vis (ACCV)*, 2018.

[32] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$ -CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, 2019.

[33] X. Yang, Q. Liu, J. Yan, and A. Li, "R3DET: Refined single-stage detector with feature refinement for rotating object," *arXiv:1908.05612*, 2019.

[34] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015.

[36] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] H. Dan and G. Kevin, "Bridging nonlinearities and stochastic regularizers with Gaussian error linear units," *arXiv:1606.08415*, 2016.

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

[39] Z. Xiao, L. Qian, W. Shao, X. Tan, and K. Wang, "Axis learning for orientated objects detection in aerial images," *Remote Sensing*, vol. 12, no. 6, p. 908, 2020.

[40] Y. Jing, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," *arXiv:1706.09579*, 2017.