Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/86/20241599

SkyNet: Multi-Scale feature augmentation and diverse expert heads for UAV aerial image object detection

Yixin Chen

Tuchuang Technology Corporation, Shenzhen, China

yixin.academics@gmail.com

Abstract. Object detection of aerial images collected by UAVs is significant in UAV missions, such as agriculture, urban planning, and traffic monitoring. Aerial images also referred to as remote sensing images (RSI), and they normally have problems like low resolution, variation of object sizes, and blurred backgrounds. The commonly seen detectors lack feature fusion and refinement module to integrate and refine semantic information and shallow features. In Addition, the detector head module that is not customized and well-designed is not adaptable to feature maps with different distributions. The problems associated with these detectors will lead to insufficient feature representation and deteriorate the detector's performance. To overcome this challenge, we propose an innovative oriented object detection framework SkyNet, including our novel efficient atrous attention module (EAAM) and a mixture of expert heads module (MOEHM). The EAAM is integrated with PAFPN to refine multi-scale semantic and contextual features. The MOEHM is for adaptively aggregation decisions from different head structures. Compared to the baseline model, SkyNet demonstrates a 0.87% increase of mAP on the DOTA dataset and a 1.2% increase of mAP12 on the HRSC2016 datasets. These results demonstrate the remarkable performance of SkyNet in oriented object detection of RSI.

Keywords: deep learning, oriented object detection, feature enhancement, remote sensing.

1. Introduction

Unmanned aerial vehicle (UAV) demonstrates extensive application and research in tasks such as urban planning, traffic monitoring, and video capturing. Oriented object detection (OOB) in RSI collected by UAVs' optical sensors includes several challenges. RSI normally includes objects with arbitrary orientations and shapes, and they might be captured in severe weather conditions. General Detectors, such as the YOLO series [1-3] utilize a horizontal bounding box (HBB) for object localization. However, because RSI includes objects with various scales and orientations, HBB usually covers regions not parts of the object. To solve this issue, OOB applies the oriented bounding box that includes angles for regression. Therefore, utilizing convolutional neural networks for OOB will significantly enhance the detection accuracy.

Two distinct modules are introduced: The efficient atrous attention module EAAM and the decision aggregation module MOEHM. Multi-scale feature fusion, assessment and decision aggregation efficiently enhance a detector's performance. Multi-scale feature enhancement generally includes fusing deeper-level semantic information with shallower-level contextual information and placing a feature enhancement module within the feature fusion module. RSI often includes small objects with various

DOI: 10.54254/2755-2721/86/20241599

shapes and orientations, posing significant challenges for detectors. The detectors commonly utilize FPN [4] after a feature extraction module for multi-scale feature enrichment and fusion. Decision aggregation means aggregating the decision from the detector's head, which predicts the classification and localization result. We used the mixture of experts training techniques for large language models to enhance feature representation while introducing minimal computational cost to aggregate the decision from different head structure structures adaptively.

2. Oriented Object Detection of Remote Sensing Images with Deep Learning

2.1. Oriented Object Detection of Remote Sensing Images

Oriented Object Detection (OOD) provides significant insight into RSI because it can detect objects with various orientations and shapes. RRPN [5] utilizes angled anchors to produce rotated proposals that align with the object's orientations. The ROI Transformer [6] convert the horizontal bounding boxes into rotated ones. However, the conventional approaches always introduce extra computational costs. Improvements such as the Gliding Vertex [7] adjust the vertex's position to enhance oriented object detection, introducing less computational cost. However, although these detectors introduce angled predictions for RSI, more robust OOD frameworks are still needed to boost detection performance.

2.2. Attention Mechanism

The Attention Mechanism typically highlights the crucial aspects of feature maps, spatial-wise or channel-wise. The Squeeze-and-Excitation (SENet) [8] is a form of channel attention that emphasizes the significant channel of feature maps. The convolutional block attention module (CBAM) [9] has channel-wise attention followed by spatial-wise attention, enhancing the features by sequentially focusing on essential channels and regions. Many two-stage detectors use the CBAM to perform feature refinement during the feature extraction stage. In order to highlight the crucial channels, the ECANet [10] utilizes the one-dimensional convolutional kernel to learn local cross-channel interaction. The SENet, CBAM, and ECANet are generally used as feature selection and enhancement techniques within the feature extraction stage.

2.3. Multi-scale Feature Enhancement and Fusion

The convolutional neural networks comprise several layers; the deeper layers generally have more semantic information, and the shallower layers have more contextual information than the deeper layers. Feature enhancement and fusion means having a feature selection or enhancement module used to refine multi-scale feature representations and fusing deeper-level features with shallower-level features. The FPN introduces a top-down pathway to aggregate semantic information to contextual information, enriching the feature representations from the CNN backbone. PANet [11] introduces a bottom-up and top-down path to further aggregate features from the CNN backbone. NAS-FPN [12] uses architectural search to configure the multi-scale feature fusion. Nonetheless, these feature fusion schemes are effective for natural images, but they lack customization for RSI due to the inherent challenging nature of RSI.

2.4. Mixture of Experts

The Mixture of Experts (MoE) is an adaptive ensemble technique, and it is widely used in natural language processing, particularly for the training of large-language models. The approaches divide the networks into submodules, each called “expert”. A gating mechanism will decide which experts will be used in learning and aggregate the results from each expert. For each input feature representation, only a few experts will be selected, and the selected experts are chosen according to the top-k selection algorithm, which will return the indices of experts with the highest gating scores. The selected experts will aggregate their weighted predictions and generate the final prediction result. The Switch Transformer [13] introduces more parameters to enrich feature representations with negligible computational cost.

DOI: 10.54254/2755-2721/86/20241599

3. Methodology

3.1. Overall Pipeline

Figure 1 demonstrates the overall architecture of the SkyNet. SkyNet comprises the feature extraction module ResNet, a multi-scale feature fusion module PAFPN, our proposed feature selection and enhancement module EAAM, the proposal network Oriented RPN, and our proposed decision aggregation module MOEHM. Our proposed EAAM are integrated with PAFPN for multi-scale feature fusion and enhancement, and our proposed MOEHM module incorporates a mixture of experts to aggregate the predictions from different expert structures adaptively.

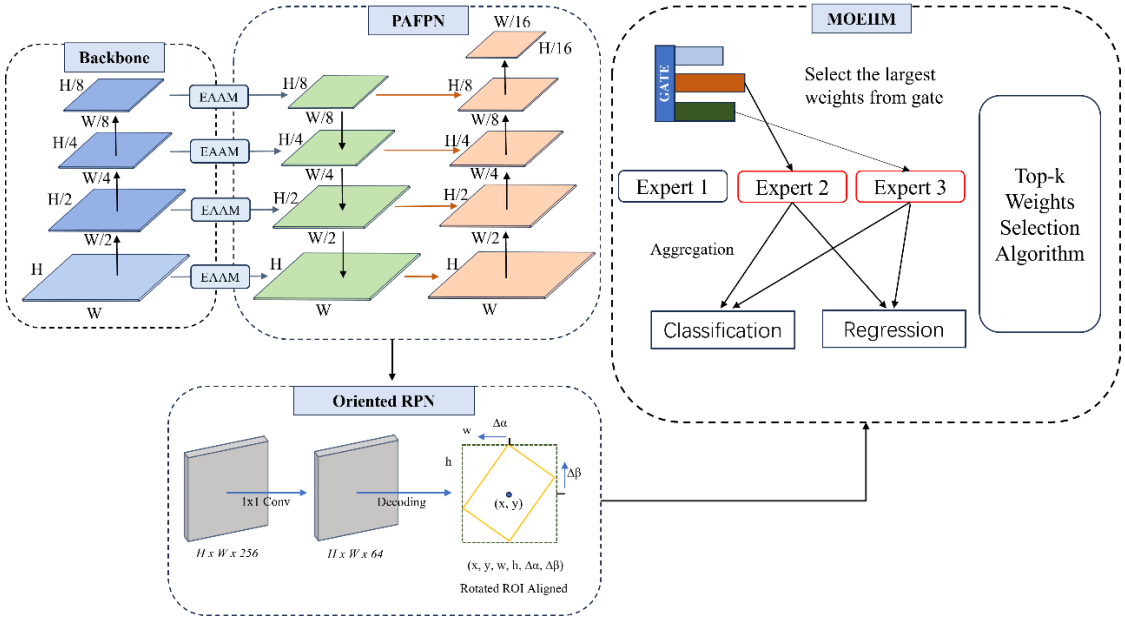

Figure 1. The overall architecture of SkyNet

We propose the EAAM for feature enhancement and refinement, which is utilized after the feature extraction stage of Resnet. The EAAM focus on essential channels and help to reduce noise and irrelevant information from each stage of Resnet. The baseline module utilizes FPN for feature fusion, but we decided to use PAFPN for enhanced feature fusion since PAFPN introduces another bottom-up path. For the MOEHM, the input feature maps from PAFPN will be fed into the gate unit, and the gate will generate scores for each expert structure we proposed. A top-k selection mechanism will select the expert for prediction based on the generated scores.

3.2. Efficient Atrous Attention Mechanism

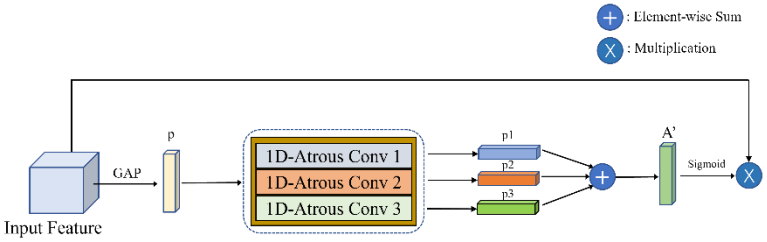

Figure 2. The structure our proposed feature refinement module EAAM

Figure 2 shows the overall structure of the EAAM. The input will be average pooled and generate the feature descriptor p, and the p will go through one-dimensional atrous convolution with different atrous

DOI: 10.54254/2755-2721/86/20241599

rates. The atrous convolution will enhance the receptive field, so they will learn different local cross-channel interactions and generate p1, p2, and p3, which will be fused and become the enriched feature representation A'. The entire workflow of the EAAM can be expressed in Equation 1, where p = GAP(Input) and σ denotes the sigmoid function.

$$Result\ Feature\ Map = \sigma\big(\text{AtrousConv1(p)} + \text{AtrousConv2(p)} + \text{AtrousConv3(p)}\big) \cdot \text{Input} \quad (1)$$

We use the atrous convolution to generate multi-scale feature descriptors to enhance the receptive field for the input sequence. The atrous convolution can increase the receptive field of the convolutional kernels without introducing extra parameters, which is achieved by inserting zeros between the filter weights. Figure 3 visualizes the one-dimensional and atrous convolution kernels with an atrous rate equal to 2. The red boxes in Figure 3 denote the regions of the standard convolution kernel and the regions of the atrous convolutional kernel. We utilize standard convolution and atrous convolution with atrous rates equal to 2 and 3 in our attention mechanism, and this setting generates optimal results.

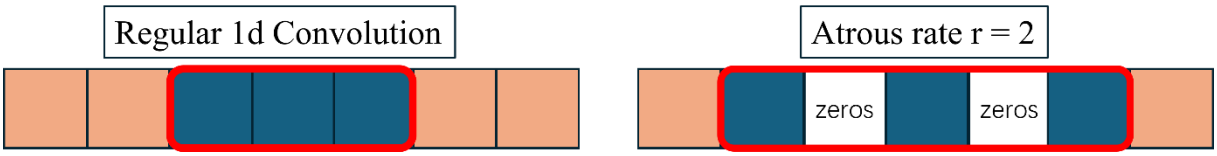

Figure 3. The visualization of one-dimensional standard convolution and atrous convolution

3.3. Mixture of Experts Head Module

Our proposed mixture of detector head modules aggregates the classification and localization results from different detector head "expert". These experts have different structures; two are single-branched, and one is double-branched. A double-branch head structure uses a convolutional head for regression and linear layers for classification. The logic of a mixture of experts starts from input to experts, which can be formulated as Equation 2, where y_i denotes the output of the i^{th} expert.

$$y_i = f_i(x) \quad (2)$$

Equation 3 illustrates the gating network computes the gating values g_i for each expert based on the input x, and W_g and b_g denote the weights and biases. In our experiment, we set the bias to zero. After the gating scores are generated, the Softmax function ensures the gating values sum to one.

$$g_i = \frac{\exp\{(W_g \cdot x + b_g)\}}{\sum_{\{j=1\}}^{\{N\}} \exp\{(W_g \cdot x + b_g)\}} \quad (3)$$

Equation 4 means that each expert's prediction y_i is aggregated by its gating value g_i, generating the weighted output z_i.

$$z_i = g_i \cdot y_i \quad (4)$$

The final output y is the weighted sum of each expert's prediction y_i, and the final aggregated result is formulated in Equation 5:

$$y = \sum_{i=1}^{N} z_i = \sum_{i=1}^{N} g_i \cdot y_i \quad (5)$$

In our experiment, we utilize two single-branch experts and one double-branch expert. A Single branch means only one branch of fully connected or convolution layers for the head module. Double-branch means we have one branch full of convolutions for regression and fully connected layers for classification. Because the input feature maps learned by previous stages are of various distributions, as remote sensing images often contain diverse and complex patterns, applying a mixture of experts allows for more customization and specialization. In our design, the model will learn from the distribution of

DOI: 10.54254/2755-2721/86/20241599

input feature maps and adaptively adjust weights for selected experts to focus on the most relevant features, enhancing overall performance and accuracy.

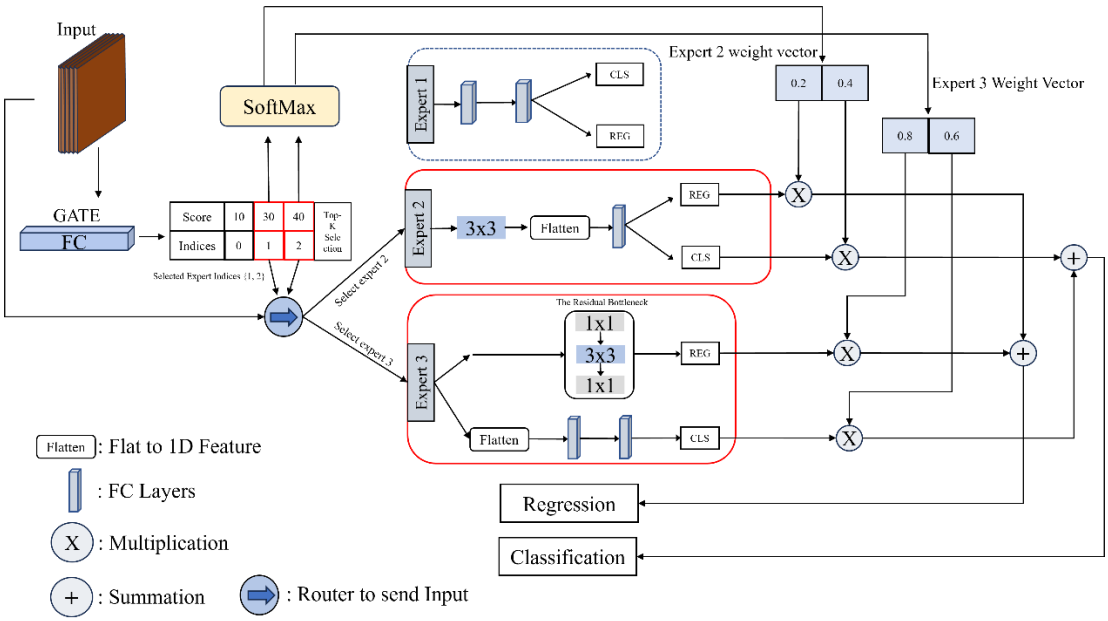

Figure 4. The structure our proposed decision aggregation module MOEHM

Figure 4 visualizes the overall design of MOEHM. The input will be fed into the gate unit, generating scores and indices for each expert. The top-k selection algorithm will select experts with the highest scores, and only the selected experts will be used to generate results, and each result will be aggregated by weights learned from the gate unit. The aggregated results will be the detector's eventual classification and regression results. Because of the memory restrictions of our GPU, we have 3 experts in our final design, and the active experts for each input feature map are 2. These experimental settings generate the best detection performance.

4. Experimental Results

4.1. Datasets

DOTA-v1.0 [14] datasets are a large-scale remote sensing benchmark comprising 2806 remote sensing images. The abbreviations of categories are shown as follows: Ground track field (GTF), Soccer-ball field (SBF), Plane (PL), Tennis court (TC), Harbor (HA), Small vehicle (SV), Baseball diamond (BD), Swimming pool (SP), Helicopter (HC), Ship (SH), Basketball court (BC), Roundabout (RA), Large vehicle (LV), Bridge (BR), and Storage tank (ST). HRSC2016 [15] is a dataset that includes only one category: ship. This dataset features six harbors worldwide and is commonly used for object detection for remote sensing imagery.

4.2. Evaluation Metrics

This research uses the mean average precision (mAP) to demonstrate the model performance. The mAP measures the accuracy of detection results, and it is the mean of the average precision (AP) score.

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{6}$$

DOI: 10.54254/2755-2721/86/20241599

In Equation 6, C denotes the total number of categories, and AP_c denotes the average precision for each category. The AP measures the precision-recall tradeoff across different threshold values and is a critical metric for detection tasks.

4.3. Comparison with State-of-the-Art Detectors

In this Section, we will demonstrate the result of our proposed framework by comparing our detectors with the previous State-of-the-Art Detectors. All models were trained and tested under uniform conditions, applying the same parameter settings to ensure fairness. The complete comparison result for the DOTA dataset is shown in Table 1. Almost for each category of DOTA, our detector shows improvements compared to the baseline model ORCNN. Our model achieves a 0.87% increment for mAP. We compared our result with ROI-Transformer, DRN, SRCDet,

Table 1. Comparison of our proposed method with other state-of-the-art detectors on DOTA.

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
ROITrans [6]	R-101	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
DRN [16]	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
SCRDeT [17]	R-101	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
R^4Det [18]	R-152	88.96	85.42	52.91	73.84	74.86	81.52	80.29	90.79	86.95	85.25	64.05	60.93	69.00	70.55	67.76	75.84
ORCNN	R-50	89.55	82.56	54.65	73.32	78.88	83.53	88.06	90.89	86.28	83.60	59.41	66.12	74.67	69.44	55.69	75.82
Ours	R-50	89.62	83.74	54.39	74.20	78.97	83.07	88.23	90.90	86.67	85.75	62.16	68.47	74.68	69.38	60.10	76.69

For the HRSC2016 dataset, Table 2 shows the result of each SOTA detector, and we used mAP07 and mAP12 as the evaluation metrics. Based on the mAP07 and mAP12 results, we can see that our proposed detectors demonstrate better detection performance. For both tables, the backbone column denotes the feature extraction network for each detector, and we only used Resnet, which contains 50 layers. Our proposed detectors significantly enhance the detection accuracy and use fewer layers and parameters, indicating our overall framework's effectiveness and our proposed EAAM and MOHEM modules.

Table 2. Comparison of our proposed method with other state-of-the-art detectors on HRSC2016.

Method	Backbone	mAP07	mAP12
ROI Trans	R-101	86.20	-
Rotated RPN [19]	R-101	79.08	85.64
R3Det [20]	R-101	89.26	96.01
ORCNN	R-50	90.36	96.40
Ours	R-50	90.60	97.60

In this Section, we will demonstrate the result of our proposed framework by comparing our detectors with the previous State-of-the-Art Detectors. All models were trained and tested under uniform conditions, applying the same parameter settings to ensure fairness. The complete comparison result for the DOTA dataset is shown in Table 1. For most categories of DOTA, our detector shows improvements compared to the baseline model ORCNN.

5. Conclusion

In this study, We propose our novel framework, Sky-Net, for oriented object detection of aerial and remote sensing images collected by the optical sensors of UAVs. Compared to the baseline ORCNN model and other state-of-the-art models, our model framework improves the model performance and maintain the computational speed. Sky-Net utilizes EAAM with PAFPN for feature selection enhancement and refinement, and it also utilizes the adaptive and sparsely gated MOEHM for adaptive decision aggregation based on the input data distribution. We validate our framework on two popular public benchmark datasets and demonstrate that our model achieved enhanced performance. This conclusion affirms that integrating feature enhancement and decision aggregation modules will be highly effective for oriented object detection of remote sensing images.

DOI: 10.54254/2755-2721/86/20241599

References

[1] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 7263-7271).

[2] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 779-788).

[3] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. 2018 Apr 8.

[4] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 2117-2125).

[5] Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X. Arbitrary-oriented scene text detection via rotation proposals. IEEE transactions on multimedia. 2018 Mar 23;20(11):3111-22.

[6] Ding J, Xue N, Long Y, Xia GS, Lu Q. Learning RoI transformer for oriented object detection in aerial images. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 2849-2858).

[7] Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia GS, Bai X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE transactions on pattern analysis and machine intelligence. 2020 Feb 18;43(4):1452-9.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 7132-7141).

[9] Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV) 2018 (pp. 3-19).

[10] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 11534-11542).

[11] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 8759-8768).

[12] Ghiasi G, Lin TY, Le QV. Nas-fpn: Learning scalable feature pyramid architecture for object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 7036-7045).

[13] Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research. 2022;23(120):1-39.

[14] Xia GS, Bai X, Ding J, Zhu Z, Belongie S, Luo J, Datcu M, Pelillo M, Zhang L. DOTA: A large-scale dataset for object detection in aerial images. Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 3974-3983).

[15] Liu Z, Yuan L, Weng L, Yang Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. International conference on pattern recognition applications and methods 2017 Feb 24 (Vol. 2, pp. 324-331). SciTePress.

[16] Pan X, Ren Y, Sheng K, Dong W, Yuan H, Guo X, Ma C, Xu C. Dynamic refinement network for oriented and densely packed object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 11207-11216).

[17] Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, Sun X, Fu K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. Proceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 8232-8241).

[18] Sun P, Zheng Y, Zhou Z, Xu W, Ren Q. R4 Det: Refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images. Image and Vision Computing. 2020 Nov 1;103:104036.

[19] Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X. Arbitrary-oriented scene text detection via rotation proposals. IEEE transactions on multimedia. 2018 Mar 23;20(11):3111-22.

DOI: 10.54254/2755-2721/86/20241599

[20] Yang X, Yan J, Feng Z, He T. R3det: Refined single-stage detector with feature refinement for rotating object. Proceedings of the AAAI conference on artificial intelligence 2021 May 18 (Vol. 35, No. 4, pp. 3163-3171).