
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 2, pp. 883–890, DOI 10.12694/scpe.v25i2.2538

LIGHTWEIGHT SALIENCY TARGET INTELLIGENT DETECTION BASED ON MULTI-SCALE FEATURE ADAPTIVE FUSION

MUQING ZHU*

Abstract. In order to solve the problems of small targets, variable shooting angles, and heights in drone images, the author proposes an adaptive drone target intelligent detection algorithm based on multi-scale feature fusion. The results show that after adding a deconvolution cascade structure to the network, mAP increased by about 2.5 percentage points and AP^{50} increased by about 3 percentage points. Compared with Method 3, Method 4 uses GA-RPN instead of RPN, and when the IOU is 75, the AP increases by 3.5 percentage points, reflecting that the target prediction candidate boxes generated using semantic features adaptively match better than the manually designed target candidate boxes. This indicates that the proposed target detection framework has better classification ability and higher frame regression accuracy. Multi scale adaptive candidate regions are used to generate fused features of different scales generated by the network, weighted fused multi-scale features are used for target prediction, and semantic features are used to guide the network to adaptively generate target candidate frames, greatly enhancing the feature expression ability of various targets and improving the detection accuracy of aerial targets.

Key words: Target detection, Deep network, Feature fusion, Multi-scale feature adaptation, Lightweight residual network model

1. Introduction. Over the past two decades, the computer vision field has shifted its emphasis from traditional approaches to an increasing focus on vision-based learning. The evolution of research in this field has seen a notable transition from conventional methods to the prominent use of deep learning techniques. The application of deep learning has gained significant importance in addressing various challenges within the realm of computer vision [12]. Productivity research is an important aspect of computer vision. It is the basis of high resolution and high level of clarity (such as separate images), image processing, scene understanding, object detection, image description, etc.The core task of target detection is to label the target concerned by the task and the target location information in the image to be detected. Since the 21st century, the performance of computing devices has been greatly improved, and the field of target detection relying on high-performance computing devices has also ushered in considerable development [8]. At the same time, the demand for intelligent industrial production continues to expand, and the market demand for target detection technology is also growing, research in the field of target detection has already begun. Currently, in-depth research-based target detection technology has been applied to autonomous driving systems; face recognition; medical images; security; fault diagnosis; military and other field [3]. The object detection method identifies the most attractive targets from the input image and is the initial step of a multi-vision computer [4]. When considering the evolution of saliency target detection, it can be categorized into two distinct approaches: traditional methods that rely on manually crafted features and heuristic priors, and task-oriented saliency target detection methods built upon deep learning. Traditional saliency target detection primarily depends on specific features like color, texture, and image gradients to compute target significance. While these methods are capable of identifying important elements in an image, they are constrained by the need for extensive prior data on significance, which can limit their effectiveness in complex environments. Traditional detection methods have low detection efficiency and long detection time. The saliency target detection algorithm based on deep learning benefits from the rapid development of Full Convolution Network (FCN), and its performance is far better than the traditional methods. FCN has powerful feature extraction ability, which can obtain edge details, texture clues, context features and high-level semantic information with multi-layer and multi-scale. However, as the number of network layers is stacked, pooling operation brings high-level abstract semantic expression, and at the same time, it also leads to image size reduction, thus losing a lot of detail information. In recent years,

*Guangzhou Huali College, Guangzhou, Guangdong, 511325 (Corresponding author, MuqingZhu9@163.com)

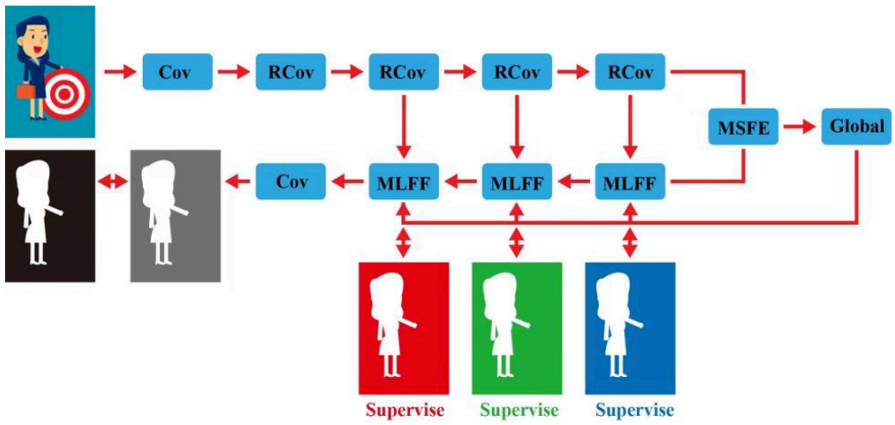

Fig. 1.1: Network framework of multi-scale feature extraction and multi-level feature fusion

a large number of end-to-end saliency detection frameworks have a large number of parameter redundancy in the encoding and decoding stages, resulting in a decline in the test speed, which cannot meet the real-time requirements. Convolution neural network can extract different levels of features, among which the advanced features have semantic features and can locate significant objects; The lower level features have rich details and can be used to sharpen the edges of prominent objects. A direct idea is to simply splice different levels of features. However, it is difficult to achieve the desired effect and capture clearer details in this way. In order to solve the above problems, in order to improve the recognition accuracy of small and medium-sized targets in low performance UAV aerial imagery, this paper presents a modified UAV The goal is to find problems based on multiple combinations. As shown in Figure 1.1, the network framework for multi-scale feature extraction and multi-level feature fusion [11].

2. Adaptive UAV Target Detection Algorithm Based on Multi-scale Feature Fusion. To enhance the precision of detecting small objects in less efficient aerial images captured by UAVs, we've introduced an adaptive algorithm for UAV object detection. This algorithm is grounded in the fusion of multiple features, ensuring improved accuracy in recognizing these small targets [5]. The first part is the Light Source Deep network (LResnet) for content extraction. Combining the advantages of residual learning, the ordinary solving tasks are divided into the deep network processing and the fast solution process, which improves the efficiency of the network. The second part is a multi-scale adaptive candidate area generation network, select the last layer C2, C3, C4, C5 of the same size output feature map generated by each layer from the four layers of LResnet, and leverage 1×1 The length of each solution is fixed at 256 by the solution. The cascade deconvolution model is employed to enhance the resolution of deep-layer feature maps. This, in turn, enables the utilization of the broader contextual information from upper-layer feature maps. Weighted fusion is applied to these maps, taking into account their channel sizes. This process results in the extraction of four distinct feature points, denoted as P2, P3, P4, P5, which exhibit robust orientation cues for target detection [6]. On each layer of features generated by deconvolution cascade network, GA-RPN (Guided Anchoring Region ProposalNetwork) is used to adaptively generate candidate boxes and corresponding category probability values for prediction targets according to semantic features, and the final prediction results are obtained through non maximum suppression [2].

2.1. Lightweight residual network model. Deeper degree and smaller receiver area of solving neural network can improve the accuracy of distribution network. The traditional solution function is to add channel parameters and convolution kernel of the input process after the solution, and the output is used as the content of the next layer. However, with the increase of the network depth, the traditional algorithms and numerical algorithms increase with the depth of the network layer, which leads to the increase of the model size, which is difficult to implement for UAV platforms with limited resources. In order to solve this problem, the depth

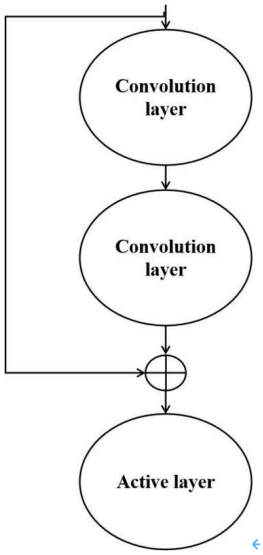

Fig. 2.1: Schematic diagram of residual structure

separation algorithm was used to optimize the model solution, that is, the model solution was decomposed into the depth and the algorithm. In the depth resolution, each channel of the input image is provided with a resolution kernel, and each resolution kernel is only responsible for the resolution of the channel image. Then use 1×1, compress and combine the output results of each channel obtained through depth convolution [10].

After the ordinary convolution is decomposed into depth convolution and point convolution, although the calculation of network model parameters is effectively reduced, the number of network layers is greatly deepened, and the gradient is easy to disappear during network training, leading to greater difficulty in model training, that is, "network regression" phenomenon. The use of residual structure can greatly reduce the burden of deep network training, connecting the shallow network with the deep network through SkipConnection is equivalent to fusing the underlying feature information into the high-level, which can ensure that the input feature information is not lost, and enhance the expression ability of features to the target, and the gradient can be well transferred to the shallow layer. In conclusion, a lightweight depth residual network model (LResnet) is constructed to extract the convolution features of aerial images. In the network structure parameters of the model, OutputSize represents the output feature size, Kernel represents the convolution kernel size, and OutputChannels represents the dimension of the output feature [7].

The residual network (ResNet) is a new deep learning network, which can reach 152 layers of convolution at the deepest. To address the issue of overfitting in deep neural networks, the author introduced a "residual" structure designed to mitigate the problem of vanishing gradients. This structure is illustrated in Figure 2.1.

The activation function is to solve the problem that the linear model in the neural network is not strong enough to add nonlinear factors. Through this function, the features are retained, the redundancy in some data is removed, and finally mapped out. The activation functions of neural network include linear function, threshold function, sigmoid function, bipolar S-shape function, hyperbolic tangent function, and now commonly used ReLU (corrected linear units) function. Apply certain mathematical and physical principles to achieve the desired effect. The network activation function used in this paper is the nonlinear PReLU function (equation (2.1)). Parameters of the negative part in PReLU α It is learnable rather than fixed. It is only necessary to give him an initial value during training, and then it can be continuously revised according to the depth of training. Compared with ReLU, PReLU introduces additional parameters, so there is no need to worry about over-fitting, and the convergence speed is faster.

$$f(x) = \begin{cases} x & x > 0 \\ \alpha(e^x - 1) & x \leq 0 \end{cases} \tag{2.1}$$

Table 2.1: Parameters of each deconvolution layer

Layer	Type	K ernel	Stride	Output size
h1	Deconvolution	3×3	1	14×14×256
h2	Deconvolution	3×3	1	28×28×256
h3	Deconvolution	3×3	1	56×56×256

2.2. Multi scale adaptive candidate area generation network. The aerial image of UAV is very large, which leads to the small proportion of targets in aerial image and the insufficient analysis content. When using depth convolution neural network to extract the image's objective features, the resolution neural network has high acceptance space and strong data representation ability, but geometric data representation ability of weak features, which is not good for small target detection; The fundamental network has a compact size, excelling at capturing geometric information while falling short in representing semantic information effectively. In addition, the traditional target detection network uses manually designed fixed size candidate boxes, different size candidate boxes need to be designed for different detection problems, and the size of the target varies greatly, the fixed design will hinder the improvement of detection accuracy. In order to solve the above problems, a multi-scale adaptive candidate region generation network is constructed based on the LResnet framework. High level semantic features are weighted into the low level feature map through deconvolution cascade structure. It enhances the expression ability of features on the target, and uses multi-level and different scale feature maps for target prediction, on each layer of features generated by the deconvolution cascade network, the position and shape of candidate frames are predicted according to image features, and sparse and arbitrarily shaped candidate frames are generated. Parameters of each layer of deconvolution cascade structure are shown in Table 2.1.

2.3. Multi-task loss function. Because the candidate region is generated adaptively, the anchor location loss function L_{loc} and anchor shape loss function are added on the basis of the traditional classification loss and regression loss. The total target loss function can be expressed as formula (2.2)

$$L = L_{\text{cls}} + L_{\text{reg}} + \beta_1 L_{\text{loc}} + \beta_2 L_{\text{shape}} \tag{2.2}$$

where β_1 and β_2 are the weighting coefficients of the multi-task loss function, with values of 1 and 0.1 respectively.

3. Results and Analysis. The experimental platform adopts i7-7700 processor, NVIDIAGTX1080Ti graphics card, 16G memory, and Ubuntu 16.04 operating system [9]. The experimental data used by the author is from the target detection data set of VisDrone UAV, including urban, rural, park, road and other natural scenes, it is obtained by the UAV platform in different positions and at different heights. There are 10 predefined categories marked in the data set, namely pedestrians, people, cars, trucks, buses, trucks, motorcycles, bicycles, tricycles with sheds and tricycles (where "pedestrians" refer to people with standing or walking postures, and "people" refer to people with other postures). Because the flight altitude and camera direction of the UAV are constantly changing, most of the target scales and shooting angles in this data set change greatly, and small targets account for a large proportion, in some data, targets are densely distributed.

3.1. Qualitative result analysis of target detection. In order to verify the effectiveness of the proposed solution in scheduling operations, the actual application scene image that is difficult to detect the UAV aerial target was extracted for testing, and the observed results of the algorithm are analyzed and shown. The algorithm in this paper can detect most targets. The results show that the algorithm uses multi-level competition between candidates in the electric field to make multi-level integration of low-level and high-level features, and use the elements semantic language to guide the network to create competitive plans, which improves the detection accuracy of small-scale aerial targets. When faced with a large number of objects and occlusion, the algorithm can also have good detection results, which shows that the algorithm in this paper uses various features to predict objects, and greatly improves the ability of feature extraction. In addition, the algorithm is very little affected by light changes, and it can still have good detection performance in the dark

Table 3.1: Comparison of feature extraction networks

Model	Size /MB	Ratio /%	Accuracy /%
Resnet	97.7	-	81.3
L Resnet	10.2	10.4	80.6

Table 3.2: Effectiveness comparison test of algorithm modules

Method	mAP	AP^{50}	AP^{75}
1Faster-RCN(Resnet50+ RPN)	18.63	35.87	17.86
2L Resnet+RPN	18.52	35.75	17.44
3LResnet+DC+ RPN	21.03	38.46	18.03
4LResnet+DC+ GA-R PN(ours)	22.1 2	38.76	21.53

environment illuminated by night lights, this also proves that the proposed algorithm can cope with a variety of environmental changes, meet the needs of actual tasks, and has a good generalization ability.

3.2. Algorithm feasibility verification analysis. To evaluate the feature extraction network's performance, an experiment was conducted to compare LResNet with ResNet. The experimental data validated the feature extraction network's efficacy when employed by the proposed algorithm. To quickly assess the algorithm's performance, a training set comprising 13,700 images from the VOC2012 dataset was selected, along with a test set of 3,425 images, all under the same conditions. The results, presented in Table 3.1, highlight that LResNet's network model is remarkably compact, with a size of only 10.2 MB, which is roughly one-tenth the size of ResNet-50's network model. However, under the same conditions, the classification accuracy of LResent and ResNet on VOC2012 is only 0.7% different, in addition, the video memory occupation rate required by the LResNet model in the running phase is also greatly reduced, only 546 MB of video memory is needed, which is about 20% of the ResNet-50 network video memory occupation rate. This shows that LResNet greatly reduces the amount of network parameters while losing very little detection accuracy, it also greatly reduces the memory usage during algorithm running.

In order to verify the effectiveness of deconvolution cascade model (hereinafter referred to as DC module) and GA-RPN in UAV aerial target detection algorithms in this algorithm, a multi-scale comparison scheme was designed based on VisDrone target detection data, as shown in Table 3.2 and Figure 3.1. The test experiment was conducted in the VisDrone test dev dataset (1610 UAV aerial images, including various situations in the VisDrone dataset), set Faster Rcnn (Resnet50+RPN) as the baseline comparison network, and conduct quantitative analysis using the evaluation indicators of mean precision (mAP) and average precision (AP, including APs with IOUs of 0.50 and 0.75, recorded as AP^{50} and AP^{75}) [1].

By comparing Method 1 and Method 2 in Table 3.2, it can be seen that after the algorithm replaces the Resnet50 with a large number of parameters with a lightweight residual network, compared with Faster RCNN, the mAP decreases by only 0.11 percentage points, it shows that the network model optimized by depth separable convolution has little impact on the detection effect, but it greatly reduces the number of network parameters [13]. By comparing Method 2 and Method 3, it can be seen that after deconvolution cascade model is added into the network, the mAP increases by about 2.5 percent content, and the AP^{50} reaches about 3 percent content. This shows that the algorithm in this paper combines the weights in the low-level map into the high-level sequence map by using the cascade decision model, and the multi-level combination algorithm gets a more robust scheme. It uses multi-level information to predict the target. It is more suitable for multi- target weather forecast that changes with UAV flight altitude. Compared with method 3, method 4 replaces RPN with GA-RPN, when IOU is 75, the AP increases by 3.5 percentage points, which reflects that the candidate frame for target prediction generated adaptively using semantic features is more matched than the target candidate frame designed artificially, it also indicates that the target detection frame proposed by the author has better classification ability and higher frame regression accuracy.

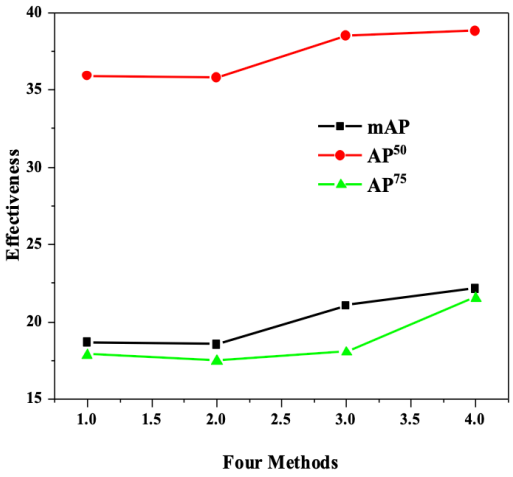

Fig. 3.1: Comparison of effectiveness of each module of algorithm

Table 3.3: Comparison of test results of various categories in VisDrone test data set

Method	Pedestrian	Person	Bicycle	Car	Van	Truck	Tricyele	Awn	Bus	Motor
Faster-RCNN	18.34	7.62	6.76	43.31	27.53	19.95	10.13	7.65	36.87	8.79
Ours	22.43	7.61	8.56	50.18	34.63	24.34	14.1 1	9.08	36.25	14.88

The detection results of the algorithm for each category in the Vis Drone test dataset are shown in Table 3.3, the comparison index is the AP of various targets. As can be seen from the Table, because Faster RCNN algorithm uses single depth feature and manually designed candidate box to predict targets, it can not adapt well to the actual situation of variable target scales and many small targets in UAV aerial photography data set, so it has poor detection effect for most types of targets. Compared with Faster RCNN, the APs of other categories except "people" and "buses" proposed by the author have improved by 1∼7 percentage points. Among them, the average detection accuracy of "pedestrian", "bicycle" and "motorcycle", which occupy a small proportion in the image, has been significantly improved, indicating that the proposed algorithm uses multi-scale adaptive candidate regions to generate fusion features of different scales generated by the network, at the same time, multiple scale features after weighted fusion are used for target prediction, and semantic features are used to guide the network to adaptively generate target candidate boxes, which greatly enhances the feature expression ability of various targets and improves the detection accuracy of aerial targets [14].

3.3. Comparison experiment of mainstream UAV target detection algorithms. During the comparative experiment phase, we assess models derived from various popular target detection networks after subjecting them to the same training data, mainly comparing the ability of different target detection algorithms to detect and recognize ground targets when the UAV is flying at low altitude, and verifying the detection performance of this algorithm. The comparison algorithm includes RetinaNet, FPN, YOLOv3 and CornerNet. In the experiment, the evaluation indexes of mAP, AP^{50}, AP^{75} and frame rate (FPS) were used to quantitatively analyze the detection accuracy and detection speed.

The results displayed in Figure 3.2 and Figure 3.3 demonstrate a significant enhancement in the accuracy of target detection indicators compared to mainstream target detection algorithms when applied to UAV aerial photography data. The mean Average Precision (mAP) has notably reached 22.12%, surpassing YOLOv3 by 1.82%. The AP^{50} serves as an effective metric to evaluate the algorithm's classification performance, while AP^{75} reflects the ability of the detection framework to precisely determine bounding box positions. Comparative experiments reveal that the algorithm proposed in this paper achieves a detection accuracy of 21.53% at an

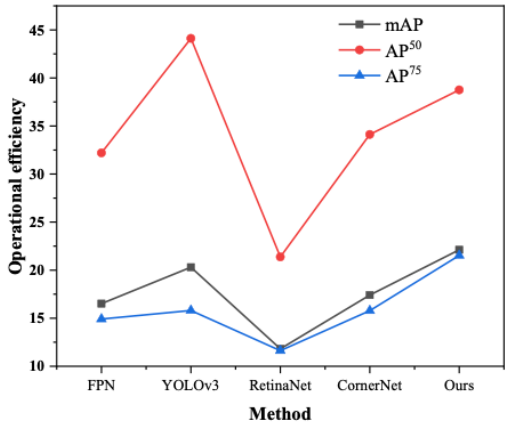

Fig. 3.2: Comparison of UAV aerial photography data of mainstream target detection algorithm (mAP, AP^{50}, AP^{75})

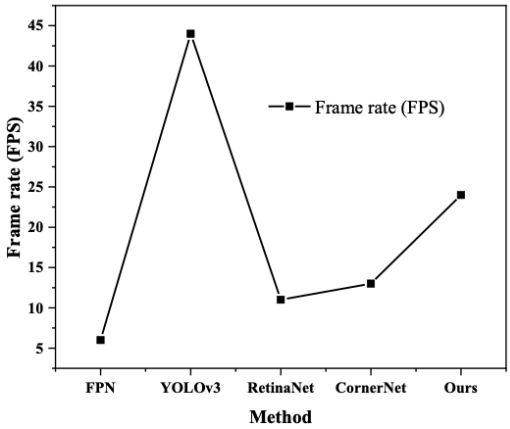

Fig. 3.3: Comparison of UAV aerial photography frame rate (FPS) of mainstream target detection algorithm

Intersection over Union (IOU) of 0.75, representing a substantial 5.73% improvement over YOLOv3. This underscores the superior classification capability and enhanced accuracy in bounding box regression of the target detection framework introduced in this paper. The notable improvement in detection accuracy can be attributed to the multi-scale adaptive candidate region generation network, which assigns weight to high-level semantic features within the low-level feature map through the deconvolution cascade structure. This approach leverages multi-level, multi-scale feature maps for target prediction, significantly enhancing the detection of diverse target types influenced by changes in UAV perspective and flight altitude. Moreover, for each feature layer produced by the deconvolution cascade network, candidate frame positions and shapes are predicted based on image features. This results in the generation of sparse and arbitrarily shaped candidate frames, which aren't constrained by fixed sizes, and therefore, they align more closely with actual target frames. This directly contributes to improved accuracy in frame regression. The detection speed of the algorithm in this paper has been significantly improved compared with the two-stage target detection algorithms FPN, RetinaNet, etc. Although it has not reached the speed level of YOLOv3, it still has the detection speed of 24frame $\cdot$ s^{-1}. The main reason is that LResnet has greatly reduced the number of network parameters and improved the operation efficiency of the algorithm.

4. Conclusion. Due to the swift advancements in science and technology, unmanned aerial vehicles (UAVs) have found extensive utility across both military and civilian sectors. UAV aerial targeting systems have become important topics in artificial intelligence and computer vision. Focusing on the problems of small measurement, high contrast and high resolution of targets in UAV aerial images, a modified UAV target detection problem is many combinations. Based on the quality of the depth separation algorithm and the residual learning, a light extraction network is created. A large number of candidate adaptive region generating network are designed, and the maps with different spatial coefficients are weighted and fused according to the channel dimensions, which improve the expression ability of the target characteristics. The key points are used to guide the network to adapt to the target parameters which match the real target better of multi- target map. The experiments showed that the algorithm successfully improves the detection accuracy of the UAV target air and has good power.

REFERENCES

[1] C. Baoyuan, L. Yitong, and S. Kun, *Research on object detection method based on ff-yolo for complex scenes*, IEEE Access, 9 (2021), pp. 127950–127960.
[2] M. Gong and Y. Shu, *Real-time detection and motion recognition of human moving objects based on deep learning and multi-scale feature fusion in video*, IEEE Access, 8 (2020), pp. 25811–25822.
[3] Z. Huo, X. Li, Y. Qiao, P. Zhou, and J. Wang, *Efficient photorealistic style transfer with multi-order image statistics*, Applied Intelligence, 52 (2022), pp. 12533–12545.
[4] P. Jia and F. Liu, *Lightweight feature enhancement network for single-shot object detection*, Sensors, 21 (2021), p. 1066.
[5] B. Li and Y. He, *A feature-extraction-based lightweight convolutional and recurrent neural networks adaptive computing model for container terminal liner handling volume forecasting*, Discrete Dynamics in Nature and Society, 2021 (2021), pp. 1–17.
[6] G. Li, X. Hao, L. Zha, and A. Chen, *An outstanding adaptive multi-feature fusion yolov3 algorithm for the small target detection in remote sensing images*, Pattern Analysis and Applications, 25 (2022), pp. 951–962.
[7] Y. Li, S. Zhang, and W.-Q. Wang, *A lightweight faster r-cnn for ship detection in sar images*, IEEE Geoscience and Remote Sensing Letters, 19 (2020), pp. 1–5.
[8] F. Qingyun, Z. Lin, and W. Zhaokui, *An efficient feature pyramid network for object detection in remote sensing imagery*, IEEE Access, 8 (2020), pp. 93058–93068.
[9] Q. Ran, Q. Wang, B. Zhao, Y. Wu, S. Pu, and Z. Li, *Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14 (2021), pp. 5786–5795.
[10] Y. Shi, J. Li, Y. Zheng, B. Xi, and Y. Li, *Hyperspectral target detection with roi feature transformation and multiscale spectral attention*, IEEE Transactions on Geoscience and Remote Sensing, 59 (2020), pp. 5071–5084.
[11] L. Wang, L. Xu, J. Shi, J. Shen, and F. Huang, *Lightweight adaptive enhanced attention network for image super-resolution*, Multimedia Tools and Applications, 81 (2022), pp. 6513–6537.
[12] L. Yanyu and L. Jinbao, *Small objects detection method based on multi-scale non-local attention network*, Journal of Frontiers of Computer Science & Technology, 14 (2020), p. 1744.
[13] J. Zhang, Y. Meng, and Z. Chen, *A small target detection method based on deep learning with considerate feature and effectively expanded sample size*, IEEE Access, 9 (2021), pp. 96559–96572.
[14] M. Zhang, Y. Chen, X. Liu, B. Lv, and J. Wang, *Adaptive anchor networks for multi-scale object detection in remote sensing images*, IEEE Access, 8 (2020), pp. 57552–57565.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Aug 25, 2023
Accepted: Oct 18, 2023