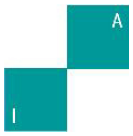
Inteligencia Artificial 24(68) (2021), 21-32
doi: 10.4114/intartif. vol24iss68pp21-32



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

A New Method of Different Neural Network Depth and Feature Map Size on Remote Sensing Small Target Detection

Yaming Cao [1, 2, A], Chen Gao [1, B], Zhen Yang [1, *]
[1] National Space Science Center, Chinese Academy of Sciences, Beijing, 100190 China.
[2] University of Chinese Academy of Sciences, Beijing, 100049 China.
[A] yafei4554@foxmail.com, [B] gaochen12@mails.ucas.edu.cn
[*] Corresponding Author: yangzhen888@sina.com

Abstract Convolutional neural networks (CNNs) have shown strong learning capabilities in computer vision tasks such as classification and detection. Especially with the introduction of excellent detection models such as YOLO (V1, V2 and V3) and Faster R-CNN, CNNs have greatly improved detection efficiency and accuracy. However, due to the special angle of view, small size, few features, and complicated background, CNNs that perform well in the ground perspective dataset, fail to reach a good detection accuracy in the remote sensing image dataset. To this end, based on the YOLO V3 model, we used feature maps of different depths as detection outputs to explore the reasons for the poor detection rate of small targets in remote sensing images by deep neural networks. We also analysed the effect of neural network depth on small target detection, and found that the excessive deep semantic information of neural network has little effect on small target detection. Finally, the verifications on the VEDAI, VEDAI-Cloud and NWPU dataset show, that the fusion of shallow feature maps with precise location information and deep feature maps with rich semantics in the CNNs can effectively improve the accuracy of small target detection in remote sensing images.

Keywords: Convolutional neural networks (CNNs); YOLO V3; Small Target Detection; Remote Sensing.

1 Introduction

In recent years, with the continuous innovation of deep learning, the performance of algorithms in speech processing, natural language processing and computer vision has been greatly improved. Object detection as an important task in computer vision has also been hugely promoted, and is widely used in military, civilian and other fields. Since the detection of small targets in remote sensing images has the difficulties of small size, few features, susceptibility to weather interference, special viewing angles, and complicated backgrounds, it is undoubtedly the most challenging research direction in the object detection, and is also a hot spot in current research [1].

In the task of target detection, there are two methods to extract features: artificial feature design and neural network feature extraction. Among them, with the continuous improvement of computer hardware computing power and the emergence of more and more datasets, the performance of deep learning method based on neural network in target detection task has reached the most advanced level. There is a lot of work to improve the detection performance in remote sensing images by modifying the existing deep network [2]. Modifications based on deep networks to adapt to remote sensing images can make full use of the powerful ability of neural network to automatically extract features. And the features extracted by different layers of the neural network are significantly different. The shallower layer can extract more information about the target texture, colour and position, while the deep layer can learn more features at the semantic level. The deeper the network, the more pooling layers, and fewer features such as the position of the small targets contained in the feature map. If only the feature maps of the last scales of the network are used for prediction, it is easy to cause the loss of small targets and missing position information [3]. Therefore, feature pyramid networks are proposed and applied to object

ISSN: 1137-3601 (print), 1988-3064 (on-line)

detection. However, for small targets in remote sensing images, the use of scale feature maps to detect common objects still does not perform well. For example, when considering the need to improve the accuracy of small target detection and make full use of the feature information extracted by Darknet 53 (YOLO V3 body), the YOLO V3 [4] network uses multi-scale feature maps with down sampling of 32 times (scale 1), 16 times (scale 2), and 8 times (scale 3) for prediction, such as Figure 1. This also means that when the size of target in the images is less than 8×8 pixels, it will be difficult to detect.

To this end, this paper has carried out further research, based on YOLO V3 to combine feature maps of different depths to detect small targets in remote sensing images. The overall network structure is shown in Figure 1. We spliced the 4 times down sampling feature maps (scale 4) from the YOLO V3 body and 2 times up sampling feature maps of scale 3 together as a new feature map to detect small targets in remote sensing images, such as Figure 1. The 4 times down-sampling feature maps does contain more small target location information. Adding new feature maps for target detection can make full use of the image information learned by the network. Feature map 5 is also added to the feature maps used for target detection, to verify whether shallower feature maps can be used to detect small targets for better results. In addition, we also considered the use of different combinations of four feature maps to carry out experiments to detect small targets in remote sensing image dataset VEDAI [5]. The fusion of shallow feature maps with precise location information and deep feature maps with rich semantics in the convolutional neural network can effectively improve the accuracy of small target detection. We also analysed the effect of neural network depth on small target detection, and found that the excessive deep semantic information of neural network has little effect on small target detection. After removing the last residual unit of YOLO V3 body, it will not change the accuracy of the network for small target detection, and will accelerate the detection speed. We compared the method in this article with SEN [6], YOLO V5 [7] and YOLO-Fine [8], and the results proved the superiority of our method.

Our main contributions are: we provided a model optimization and improvement method for the detection of small targets such as remote sensing images in the future. And a series of experiments were conducted on the question of what degree and number of feature maps were used to predict the best detection of small targets in remote sensing images, and a quantitative comparison was made to draw conclusions; the effects of the information features extracted by the last layers of neural network on small target detection were analysed experimentally.

The remainder of this paper is organized as follows. Section 2 presents the related work on remote sensing object detection. Section 3 details the method and the structure of YOLO V3, and datasets used for the experiment and the experiment setup are also introduced in this section. Section 4 shows the experimental results and brief analysis. Section 5 draws the conclusion of this paper and briefly introduces the future work plan.

2 Related Work

Before the popularization of deep learning, object detection is basically implemented by manually designing features and sliding windows. The target detection in the remote sensing image is also solved by this method. The authors of [9] first explored vehicle detection in remote sensing image by using multiple features (HOG (histogram of oriented gradients), LBP (local binary pattern), and Opponent Histogram) and the IKSVM (intersection kernel support vector machine). The authors of [10] detected the vehicle locations by a sliding window mechanism using ICFs (integral channel features) and an AdaBoost classifier in soft-cascade structure. The authors of [11] proposed the superpixel segmentation technique along with fast sparse representation to generate relevant vehicle patches. The HOG features of these patches were extracted and used in an SVM classifier for vehicle detection. The authors of [12] also proposed a catalog-based approach to detect cars in UAV (unmanned aerial vehicle) images. This method, based on the combination of artificial design features and a classifier, has developed very well in the field of object detection. However, this method consumes a lot of human resources and a large time cost when designing features. And it is difficult to design effective features for small target detection in large-scale remote sensing images.

With the development of deep learning technology, various deep learning-based algorithms have achieved gratifying results in visual tasks such as object classification, recognition, and object detection. Object detection model based on deep learning has achieved the state-of-the-art performance in different datasets in terms of accuracy, and has recently been widely used in object detection in remote sensing images. The authors of [13] have demonstrated that deep features from everyday objects generalize well to remote sensing domains. However, if these state-of-the-art models are directly applied to detect the object in remote sensing, the performance is poor due to the different characteristics of ground view images and aerial view images. The authors of [14] proposed a vehicle detection method from satellite images through HDNNs (hybrid deep convolutional neural networks).

They extracted variable-scale features by using HDNNs. The authors of [15] thought that the coarse feature maps and complex backgrounds were the main reasons why Faster R-CNN [16] does not perform well in the detection of small vehicles in remote sensing images. They adopted a HRPN (hyper region proposal network) to extract vehicle-like targets with a combination of hierarchical feature maps, and replaced the classifier after RPN (region proposal network) [16] by a cascade of boosted classifiers to reduce false detection by negative example mining. A weighted bi-directional feature pyramid network in [17] was proposed to make multi-scale feature fusion easy and fast. And a scale adaptive proposal network (SAPNet) in [18] was proposed to improve the accuracy of multiobject detection in remote sensing images. They used a final detection subnetwork in which fusion feature layer has been applied for better multiobject detection. The authors of [19] examined the applicability of object proposal methods for vehicle detection in aerial images, and overcome drawbacks of the original Fast R-CNN and Faster R-CNN for small objects as in the case of aerial images by changing the scale and number of proposals. The authors of [8] proposed YOLO-Fine to be capable of detecting small objects. In essence, this method is one of all the experiments in this paper, such as Figure 3. And this method in [8] is straightforward and it is lack of theoretical analysis.

3 Methodology

3.1 YOLO V3 Model

YOLO V3 is a single-stage deep learning detection model improved on the basis of YOLO V2 [20] and YOLO V1 [21]. This model is mainly composed of two parts: YOLO V3 body and YOLO V3 head. The body part uses Darknet 53 network structure, a total of 53 convolutional layers, consisting of 5 residual modules, each residual module consists of multiple residual units. Each residual unit consists of 2 conv2d and 1 res_block. Yolo_block is used to process different scale feature maps output by YOLO V3 body. YOLO V3 down samples the input image five times, and predicts the target after the last three down sampling. Three scale feature maps are used for target detection. Small-scale feature maps can provide high level semantic information for detection, and large-scale feature maps can provide target location information for detection. In addition, the feature maps of different scales are fused by means of up sampling, so the model has good detection effect for large-scale and small-scale targets.

However, this model performs poorly when directly used to detect small targets in remote sensing images. We think that the possible reason is that the shallow position information is not used for detection, and the deep high-dimensional semantical information after multiple down sampling has little effect on the detection of small targets. To this end, based on the YOLO V3 model, the feature maps after the first and second down sampling are used for target prediction. Experimental results prove that the improved model can better adapt to the detection of small targets in remote sensing images.

3.2 Proposed Model

The YOLO V3 model has as many as 53 convolutional layers for image feature extraction, so we believe that the reason of the model's poor performance in small target detection tasks in remote sensing images may be that shallow features are not used for detection, rather than the features extracted by the model are not sufficient. Therefore, in order to obtain more feature information of small targets and make full use of the features extracted by the YOLO V3 body, we improve the model and try to use more feature maps for detection. Our purpose is not to achieve state-of-the-art detection rate on the VEDAI and NWPU datasets, but to experiment and validate the capacity of our method to detect small objects from remote sensing images.

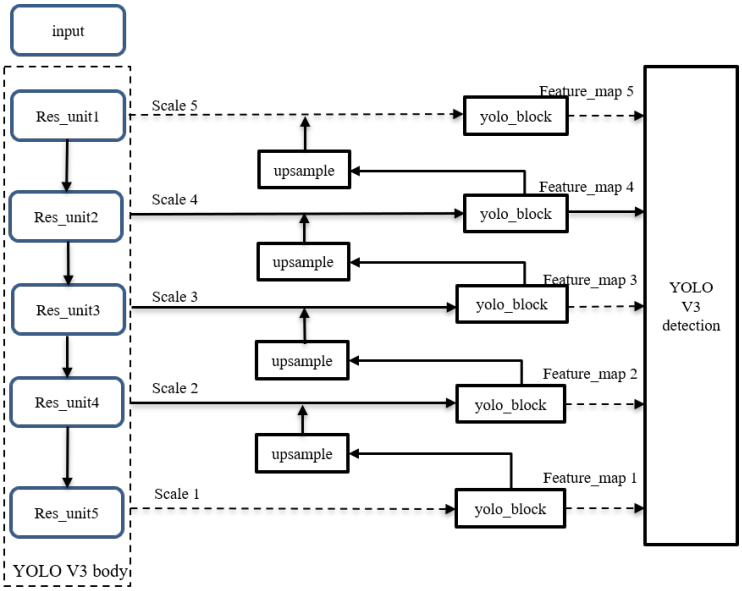


Figure 1. The YOLO V3 model with changed structure.

We try to improve the YOLO V3 model in two steps. First, keep the YOLO V3 body structure fixed, that means all models have the same output in YOLO V3 body. On the basis of the original detection using the last three down sampling feature maps (scale 1, scale 2, scale 3), the first down sampling feature maps (scale 4, scale 5) are added for detection. The 8-fold down sampling feature map (scale 3) is up sampled and stitched with the 4-fold down sampling feature map (scale 4) to obtain the fusion target detection layer (Feature_map 4). The combination of the two feature maps is achieved by the *concat* of equation (1). In equation (1), X_1 and X_2 is the two feature maps to be spliced. c is the dimension of the feature map, and K is the corresponding convolution kernel. *Concat* is to splice two feature maps in a certain dimension. The dimensions of the two feature maps in *concat* must be the same except for the spliced dimensions. Feature_map 5 is also obtained in this way. Different from the YOLO V3 model directly using Feature_map 1, Feature_map 2, and Feature_map 3 detection, we try to use different combinations of feature maps (Feature_map 1~5) to detect small targets in remote sensing images. Considering the depth of the feature map, we combined the above feature maps as follows to conduct the experiment: model_1, model_2, model_3, model_4, and model_5, such as Figure 2. From the Figure 4, we can see that model_4 achieved the best results.

$$concat\left(X_1, X_2\right) = \sum_{i=1}^{c} x_1^i * K_i + \sum_{i=1}^{c} x_2^i * K_{i+c} \tag{1}$$

$$X_4 = concat(X_5^{up}, X_4^{'}) \tag{2}$$

$$X_3 = concat(X_4^{up}, X_3^{'}) \tag{3}$$

$$Y_1=\{X_5, X_4, X_3\}=\{X_5, concat(X_5^{up}, X_4^{'}), concat(\left(concat(X_5^{up}, X_4^{'})\right)^{up}, X_3^{'})\} \tag{4}$$

$$\begin{aligned} Y_2&=\{X_2\}=\{concat(X_3^{up}, X_2^{'})\}=\{concat((concat(X_4^{up}, X_3^{'})^{up}, X_2^{'})\} \\ &=\{concat((concat((concat(X_5^{up}, X_4^{'}))^{up}, X_3^{'})^{up}, X_2^{'})\} \end{aligned} \tag{5}$$

A mathematical model is established to explain the above experimental results, and the theoretical analysis of the proposed method is carried out. X_2, X_3, X_4 and X_5 represent Feature_map 4, Feature_map 3, Feature_map 2 and Feature_map 1 respectively. Y_1 in equation (4) represents the combination of feature maps used by YOLO V3 for detection. According to equations (2) and (3), the feature maps set used for detection can be obtained, including X_3, X_4 and X_5. Among them, $X_3^{'}$ and $X_4^{'}$ represent the feature maps without *concat*. X_4^{up} and X_5^{up} represent the feature maps used for *concat* after up sampling. Y_2 in equation (5) represents the combination of feature maps used by model_4 for detection. From the expressions of Y_1 and Y_2, it can be found that the size of the feature map used for detection in model_4 is more abundant, and $X_2^{'}$ contains more small target features and information. Therefore, the experimental results of model_4 are better than those of YOLO V3.

Second, keep the detection structure of model_4 fixed, and delete the residual module in the YOLO V3 body to experiment to observe which feature layers have no effect on small target detection. That means all models have different outputs in YOLO V3 body. These models are model_4, 3scale_model_4, 2scale_model_4 and 1scale_model_4, such as Figure 2. Figure 1 shows the structure of YOLO V3 model, in which all the solid and dotted line links indicate all the structural modifications in this paper. The solid line part is the 3scale_model_4, and the corresponding relationship between the specific structure and the model is shown in Figure 2.

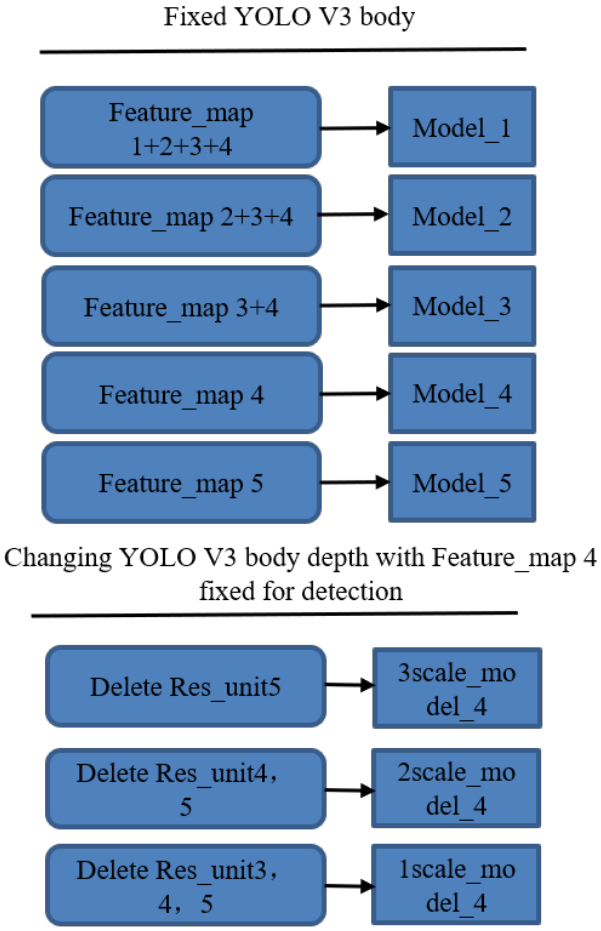


Figure 2. The way of all models composed of different feature maps and different depth YOLO V3 body.

In [8] YOLO-Fine was proposed to detect small objects in remote sensing images under different backgrounds. The essence of YOLO-Fine method is to delete the Res_unit4 and Res_unit5 on the basis of the YOLO V3 model, and use Feature_map 3+4+5 to carry out multi scale detection. This method is straightforward and it is lack of theoretical analysis and comparative experiments of various network structures. In contrast, this paper makes full comparison experiment on the combination of network depth and feature map, and makes theoretical analysis of the experimental results from the angle of the composition of the feature map used for detection. And the results on VEDAI, VEDAI-Cloud and NWPU show the superiority of the proposed model in the paper.

Figure 3. The network structure of YOLO-Fine method.

3.3 Dataset and Experimental Setup

The VEDAI (Vehicle Detection in Aerial Imagery) is a dataset of vehicle detection in remote sensing image. The dataset is formed by dividing the satellite image in a large field of view into 1024×1024 pixels and all images have been taken from the same distance to the ground. An equal number of 512×512 pixels images are obtained

after down sampling each of the above images to make the targets smaller. Therefore, there are two sizes of images in this dataset. The ground sampling distance (GSD) of the original image is 12.5cm/pixel, and the GSD of the image after down sampling is 25cm/pixel. As a dataset for benchmarking the object detection algorithm in an unrestricted environment, the dataset contains different variability in addition to very small vehicles, such as multiple directions, light and shadow changes, mirrors reflection or occlusion. In order to study the effect of the improved model on the detection of small targets in remote sensing images, this paper only uses 512×512 pixels images for experiments, and does not resize and crop the images.

The VEDAI dataset contains nine different classes of vehicles, namely the 'plane', 'boat', 'car', 'truck', 'tractor', 'camping car', 'van', 'pickup', and the 'other' category. There is an average of 5.5 vehicles per image, and they occupy about 0.7% of the total pixels of the images. There are two ways to define the size of a small target in deep learning. One is the definition of relative size. If the length and width of the target is less than 10% of the original image size, it can be considered as a small target. The other is the definition of absolute size, that is, the size of a target less than 32×32 pixels can be considered as a small target. The small target of remote sensing image in this paper refers to the target whose size conforms to the above absolute definition of small target size. The average pixel of the target in VEDAI data set is 20×20, which is undoubtedly a small remote sensing target. The number of various types of targets in the dataset is shown in Table 1.

Table 1: Statistics of VEDAI dataset [5].

Class name	Total	Orientation
Boat	170	[-π π]
Camping car	390	[0 π]
Car	1340	[-π π]
Other	200	[0 π]
Pickup	950	[-π π]
Plane	47	[-π π]
Tractor	190	[-π π]
Truck	300	[-π π]
Vans	100	[-π π]

To prove the validity of the proposed method, YOLO V3, SEN, YOLO-Fine and 3scale_model_4 are tested on the VEDAI-Cloud and NWPU [22] datasets. Compared with conventional images, remote sensing images are more vulnerable to cloud and fog occlusion and light changes. Especially considering that 66% of the earth's surface is often covered by cloud and fog [23], accurate detection of small target in remote sensing images under the interference of clouds and fog has become a problem that must be faced and solved [24,25]. Therefore, based on the VEDAI dataset, the VEDAI-Cloud dataset is constructed by artificially adding cloud interference, such as Figure 4. The NWPU dataset contains 800 high-resolution satellite images cropped form Google Earth and Vaihingen datasets and then manually annotated by experts. NWPU is a challenging 10-class geospatial object detection dataset, which can be used for both single class and multi-class objects detection.

Figure 4. Comparison of VEDAI and VEDAI-Cloud datasets.

The computer conditions for the experiment are as follows: the system is Ubuntu 16.04, and deep learning framework uses TensorFlow, and the GPU used is NVIDIA GeForce TITAN V. The size of the anchors in the newly added feature map is obtained using the K-means method. All models were initialized with COCO weights. The batch size is set to 8, and the learning rate is 0.001. In order to facilitate the follow-up researchers to compare with the experimental results of our method, after this paper is accepted, all the code in this paper will be published in GitHub.

4 Experiment

According to the improved YOLO V3 model of methodology for the detection of small targets in remote sensing images, we conducted experiments on all the changed models. A detected bounding box P and a ground truth T is considered as a correct match if:

$$\frac{Area(P \bigcap T)}{Area(P \bigcup T)} \geq 0.5 \tag{6}$$

All model test results adopt the quantitative evaluation in (6).

When the YOLO V3 body structure is fixed, we propose five models that use different combinations of feature maps for comparison with the YOLO V3 model. The test results of all models are shown in Table 2. We also tested YOLO-Fine in all datasets, and the results in VEDAI are shown in Table 2. And we can observe in Figure 5 that model_4 achieved the best detection results, compared to YOLO V3, which has improved by more than 10% on the map. This means that when the structure and output of YOLO V3 body are not changed, only using Feature_map 4 to detect small targets in the VEDAI dataset can achieve the best results. The detection results in Figure 5 show that the addition of shallow feature maps can improve the detection performance of the model on small targets, but it does not mean that the shallower the feature layer, the better the results, such as the results of model_5.

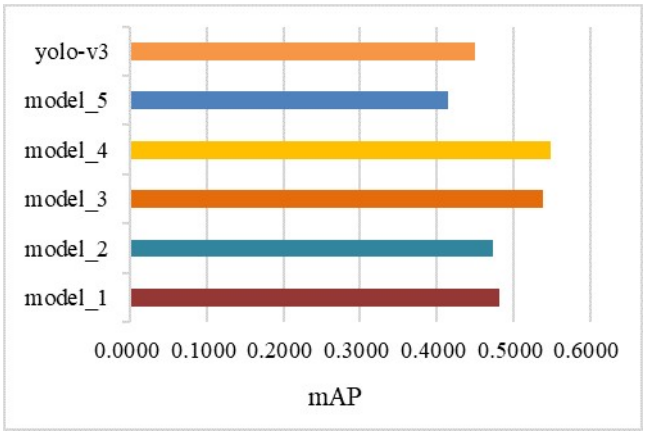


Figure 5. Detection results of the YOLO V3 body fixed and different feature maps used.

The above experiments show that the shallower feature layers may contain more position information of small targets. When these layers are used for detection, the detection results of the model can be improved. However, it can be seen from model_5 in Figure 5 that for small target detection, it is not that the shallower the feature map, the better the effect. There may be a feature map with a specific depth for the best detection effect for the current size target.

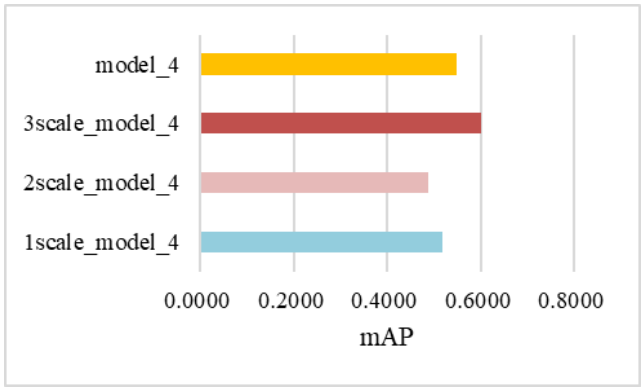


Figure 6. Detection results of the combination of different YOLO V3 body outputs and model_4.

The next experiment is to analyse the effect of the deep high-dimensional information of the feature extraction network on the detection results of small targets, such as YOLO V3 body. So, we keep the detection structure of model_4 fixed, and delete one residual module in the YOLO V3 body at a time. Four types of models are formed by combining the output of different scale of the YOLO V3 body with the model_4. The test results of these models are compared as shown in Figure 6. The 3scale_model_4 achieves best results, which shows that the last 32-fold down sampling residual module of YOLO V3 body has little effect on the detection of small targets, and even affects the detection results. And from the comparison of the test results in Figure 7, we can find that 3scale_model_4 has fewer missed detections and misjudgements than model_4. When the last residual module of the YOLO V3 body is removed, the map of the detection result is increased by 5%, and the training time is also shortened about 15 minutes.

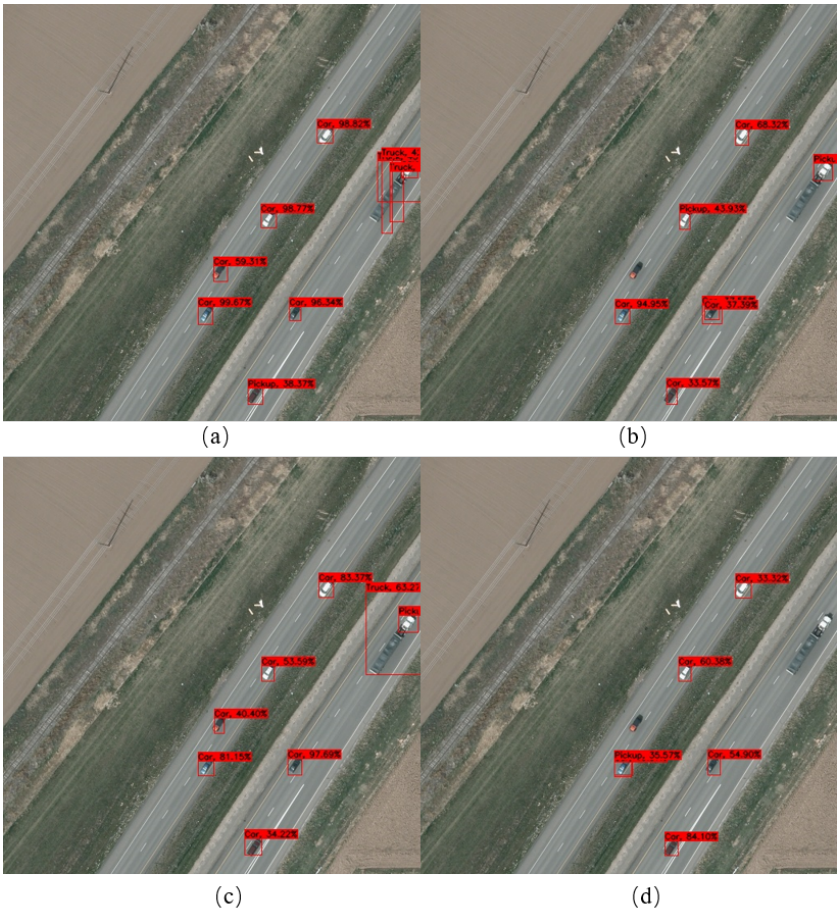


Figure 7. Detection results of different models for the same image: (a) 1scale_model_4; (b) 2scale_model_4; (c) 3scale_model_4; (d) model_4.

Table 2: Detection results of all models on VEDAI dataset.

Model	Car	Truck	Tractor	Camping car	Other	Van	Boat	Plane	Pickup	mAP	Fps
Model_1	0.7427	0.2367	0.3526	0.4911	0.4070	0.5289	0.2982	0.5305	0.7499	0.4820	17.5
Model_2	0.7005	0.2487	0.3309	0.5059	0.3017	0.4318	0.2473	0.7528	0.7444	0.4738	18.2
Model_3	0.7852	0.2649	0.3982	0.5320	0.3921	0.4954	0.2919	**1.0000**	0.6971	0.5396	18.7
Model_4	0.7934	0.2742	0.3975	0.4766	0.4029	0.5377	0.3227	**1.0000**	0.7404	0.5495	19.5
Model_5	0.7143	0.0835	0.3027	0.3989	0.2076	0.3832	0.0222	**1.0000**	0.6204	0.4148	19.3
YOLO-V3	0.6910	0.3280	0.5250	0.5470	0.3140	0.2110	0.0490	0.9320	0.4470	0.4493	19.2
YOLO-Fine	0.8353	**0.3346**	0.3895	0.6273	0.4790	0.4959	0.1815	0.9333	0.7037	0.5533	22.7
1scale_model_4	0.8344	0.2042	0.4175	0.5947	0.2716	0.3569	0.2894	**1.0000**	0.6934	0.5180	19.8
2scale_model_4	0.7818	0.1985	0.3191	0.5822	0.3102	0.4252	0.0491	**1.0000**	0.7259	0.4880	**24.9**
3scale_model_4	**0.8355**	0.2289	**0.5526**	**0.6393**	**0.4401**	**0.5872**	**0.3416**	1.0000	**0.7727**	**0.5998**	22.1

Although all target sizes conform to the definition of remote sensing small target, the final accuracy of the target is more related to the aspect ratio and geometric shape of the target than the target size. The detection accuracy of plane is very high, which is related to the unique geometric shape of plane, and it is not easy to be confused with other targets. Compared with plane, car and pickup, the accuracy of truck is very low, which is

related to the geometric shape of the truck. The length width ratio of truck is large and the boundary is difficult to distinguish.

In order to prove the effect of the model after modifying the structure, we compared mAP and Fps with SEN and YOLO-Fine. The authors of SEN increased the detection rate of YOLO V3 in small target datasets by adding new structures in the network. It can be seen that our proposed method of modifying the network structure achieved the best mAP results from the comparison in Table 3. From the comparison between the results of YOLO V3 and YOLO V5, it can be seen that for small targets detection in remote sensing images, it does not mean that the more complex the network structure, the better the experimental results. In terms of detection speed, our model is higher than YOLO V3, but lower than other state-of-the-art models, which is also means that our algorithm has optimized space in detection speed.

Table 3: Comparison results with state-of-the-art method on VEDAI dataset.

Methods	mAP/%	Time/Fps
SEN	47.8	**35.4**
YOLO-Fine	55.33	22.7
YOLO V3	44.93	19.2
YOLO V5	43.0	29.6
Model_4(Ours)	54.95	19.5
3scale_model_4(Ours)	**59.98**	22.1

All models are trained to the best on the VEDAI-Cloud and NWPU datasets, and the test results are shown in the Table 4. It can be seen that 3scale_model_4 has achieved the highest mAP on both datasets from the result comparison. The experimental results also show that the proposed method has a good effect for small target detection in remote sensing images. From the comparison of the experimental results of VEDAI and VEDAI-Cloud, we can also find that the cloud occlusion has a great interference on the remote sensing small target detection, which greatly reduces the detection accuracy of YOLO V3, YOLO V5, YOLO-Fine and SEN.

Table 4: Comparison results (mAP/%) with state-of-the-art method on VEDAI-Cloud and NWPU dataset.

Data / Model	VEDAI-Cloud	NWPU
YOLO V3	30.6	68.73
YOLO V5	34.7	72.85
YOLO-Fine	52.94	64.53
SEN	40.0	60.85
3scale_model_4(ours)	**59.94**	**77.99**

The test results of all models are shown in Table 2. 3scale_model_4 achieved the best results in terms of detection accuracy, indicating that a network with a specific depth and size for a specific size target can achieve the best detection results. According to the Fps in Table 3, reducing the network depth is not only more effective for the detection of small targets, but also improve the detection speed.

5 Conclusion

We have proposed a model improvement method that can greatly improve the effect of YOLO V3 model on small targets detection in remote sensing images. When using some models that have achieved good results in ground-level object detection to detect small targets in remote sensing images, the model can be improved by using the combination of shallower feature layer of the model to detect and modify the depth of the extracted feature network. The fusion of shallow feature maps with precise location information and deep feature maps with rich semantics in the CNNs can effectively improve the accuracy of small target detection in remote sensing images. And when the feature extraction network is deep to a certain extent, it has little effect on the detection of small

targets, and even affects the detection effect. The experimental results also prove that changing the structure of the network is effective. And it is necessary to determine the best scale and depth of the network for different size target through many experiments.

In the future work we will try to optimize the algorithm in terms of detection speed, and try to obtain the quantitative relationship between the depth and size characteristics of the model and the size of the detection target through neural network interpretability.

References

[1] Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5, 8-36, 2017.

[2] Nina, W.; Condori, W.; Machaca, V.; Villegas, J.; Castro, E. Small Ship Detection on Optical Satellite Imagery with YOLO and YOLT. *Advances in Information and Communication*: 2020.

[3] Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. *In Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117-2125, 2017.

[4] Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[5] Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery : A small target detection benchmark. *Journal of Visual Communication & Image Representation*, 34, 187-203, 2016.

[6] Ju, M.; Luo, J.; Zhang, P.; He, M.; Luo, H. A simple and efficient network for small target detection. *IEEE Access*, 7, 85771-85781, 2019.

[7] Glenn Jocher, A.S., Jirka Borovec. yolov5: v3.1 - Bug Fixes and Performance Improvements (Version v3.1). Zenodo. 2020, doi:http://doi.org/10.5281/zenodo.4154370.

[8] Pham, M.T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images. *Remote Sensing*, 12, 2501, 2020.

[9] Shao, W.; Yang, W.; Liu, G.; Liu, J. Car detection from high-resolution aerial imagery using multiple features. *In Proceedings of 2012 IEEE International Geoscience and Remote Sensing Symposium*, 4379-4382, 2012.

[10] Liu, K.; Mattyus, G. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geoscience & Remote Sensing Letters*, 12, 1938-1942, 2015.

[11] Chen, Z.; Wang, C.; Luo, H.; Wang, H.; Chen, Y.; Wen, C.; Yu, Y.; Cao, L.; Li, J. Vehicle Detection in High-Resolution Aerial Images Based on Fast Sparse Representation Classification and Multiorder Feature. *IEEE Transactions on Intelligent Transportation Systems*, 17, 2296-2309, 2016.

[12] Moranduzzo, T.; Melgani, F. Automatic car counting method for unmanned aerial vehicle images. *IEEE Transactions on Geoscience and Remote Sensing*, 52, 1635-1647, 2013.

[13] Penatti, O.A.; Nogueira, K.; Dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? *In Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 44-51, 2015.

[14] Chen, X.; Xiang, S.; Liu, C.-L.; Pan, C.-H. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, 11, 1797-1801, 2014.

[15] Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. *Sensors*, 17, 336, 2017.

[16] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *In Proceedings of Advances in neural information processing systems*, 91-99, 2017.

[17] Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. *In Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Zhang, S.; He, G.; Chen, H.B.; Jing, N.; Wang, Q. Scale Adaptive Proposal Network for Object Detection in Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 1-5, 2019.

[19] Sommer, L.W.; Schuchert, T.; Beyerer, J. Fast deep vehicle detection in aerial images. *In Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*,311-319, 2017.

[20] Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. *In Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263-7271, 2017.

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

[22] Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *Isprs Journal of Photogrammetry & Remote Sensing*, 117, 11-28, 2016.

[23] Zhang, Y.; Rossow, W.B.; Lacis, A.A.; Oinas, V.; Mishchenko, M.I. Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data. *Journal of Geophysical Research: Atmospheres*, 109, 2004.
[24] Lv, H.; Wang, Y.; Shen, Y. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. *Remote Sensing of Environment*, 179, 183-195, 2016.
[25] Li, Q.; Lu, W.; Yang, J. A hybrid thresholding algorithm for cloud detection on ground-based color images. *Journal of atmospheric and oceanic technology*, 28, 1286-1296, 2011.