Improved RetinaNet model for the application of small target detection in the aerial images

Improved RetinaNet model for the application of small target detection in the aerial images

Hong Tian[1, a], Yufu Zheng[2, b], Zhaozhao Jin[3, c]

[1,2,3]Electronics and Information Engineering College of Lanzou Jiaotong University, Lannzhou, Gansu, China

[a]email: 1012376948@qq.com, [b]email: 46328092@qq.com, [c]email: @1217602826qq.com

Abstract. The target detection algorithm for accurately locating the location of objects in the image and their categories is widely used in intelligent traffic, forest fire prevention, electric power patrol and other fields. Existing target detection algorithms detect objects in head-up images, and their accuracy is not high when used in aerial image detection. To solve this problem, a small target detection algorithm based on improved RetinaNet model is proposed. First, replace the ResNet101 deep residual network in the original network with ResNet152, Independent of image features, The model is accelerated by merging ScaleNet network structure; Besides, in order to increase the feeling of the small target detection, to upgrade the existing FPN network output, increase the P2 feature layer, improved the range makes the network overall receptive field and the robustness of small target detection. Through the experimental result shows that compared with the original RetinaNet algorithm, The performance index of the improved algorithm is obviously improved. the average detection accuracy of the improved algorithm for small target detection is quite high 6.21%, it has a good detection effect in many scenarios.

1. Introduction

With the improvement of computer hardware, the continuous development of artificial intelligence technology and different kinds of target detection technologies, the processing of aerial images has changed from traditional methods to deep learning. It is of great significance to study aerial image target detection based on deep learning for intelligent transportation [1], forest fire prevention [2], power inspection [3] and other fields. There are two types of target detection algorithms based on deep learning. The representative network of one stage models is YOLO(you only look once) [4], SSD (Single shot MultiBox Detector) [5], Yolov3 [6] and so on. The second category is a two-stage target detection algorithm. The following networks are representative models: R-CNN [7], SPP-Net [8], Fast-RCNN [9], Faster-RCNN[10], R-FCN[11], etc. Common One-Stage and Two-Stage target detection methods have their own advantages and disadvantages. The advantage of two-stage detection is that it has higher recognition accuracy and positioning accuracy for the detected object. Its disadvantages are: because the target candidate area is to be generated to prepare for the next step, the two tasks are carried out in series, so it takes a long time and can not achieve real-time detection; The advantages of one-stage detection are: fast detection speed, and satisfying scenes with high real-time requirements. The disadvantage is that when operating on the feature map of each layer, a large number and dense candidate frames will be generated, which will lead to a large number of negative samples. Compared with the two-stage detection, the recognition accuracy and positioning accuracy

have declined, and the missing detection of small objects and overlapping objects is obvious. There are great differences between aerial images and head-up images used in general target detection, which are mainly manifested in the observation of target features, the distribution of target positions in the image and special visual angles. If the detection model applied to the head-up image is directly applied to the aerial image, the detection effect will be poor because of the huge difference between the two, so it must be further improved to adapt to the characteristics of the aerial image. Literature [12] uses deformable convolution network to extract features of remote sensing image targets with scale and direction changes, and then predicts and discriminates the deformed features extracted by multi-layer residual module, which improves the detection performance of small targets in aerial images. Literature [13] proposes a target detection method for remote sensing images, which combines Convolutional Neural Network and Hybrid Boltzmann Machine, and further enhances the accuracy of target representation combined with context information, thus improving the detection accuracy of small targets in aerial images. At present, most models are used to detect aerial image targets by improving the commonly used one-stage detection and two-stage detection algorithms. Although the detection accuracy and precision have been improved, the balance between them cannot be achieved. The appearance of RetinaNet [14] algorithm has greatly improved this problem. It not only has the detection speed of a single-stage detection algorithm, but also has the same detection accuracy as the two-stage detection algorithm. As a benchmark model for aerial image target detection, it is more in line with actual needs. In order to solve the problem of low detection accuracy of small targets in aerial images, this paper takes RetinaNet as the benchmark model, increases ResNet from 101 layers to 152 layers in order to obtain more accurate original image features, replaces the existing 3×3 convolution of ResNet152 with a Scale Aggregation block to form a scale network, and adds P2 feature layer on the basis of it, which increases the receptive field range of the whole network. Comparing the differences of RetinaNet network after adjustment, and carrying out simulation test, the results show that the recall rate, detection rate, average detection accuracy, missed detection rate and other aspects of target detection after optimization and improvement of RetinaNet network are enhanced to varying degrees, and almost no additional computing resources are added, which can achieve both detection accuracy and speed.

2. Introduction of RetinaNet model

RetinaNe model is mainly composed of three parts: Residualnetwork (ResNet)[15], which is used to extract image features; A Featurepyramidnetworks (FPN)[16] for further processing features; Classification and Return subnetwork, which is used to output the final detection results of the network. The number of layers of neural network is directly proportional to the abstract degree of feature extraction and adjustable parameters. The higher the number, the better the fitting effect. However, there are problems such as gradient explosion and gradient disappearance. ResNet's residual unit structure adds a connection to the convolution feedforward network, which makes it possible to train a deeper neural network. After the image passes through deep ResNet, the features are preliminarily extracted, and then the feature images of different layers are fused by FPN bottom-up, top-down connection and horizontal connection, and finally sent to classification and return subnetwork. Its network structure is shown in Figure 1:

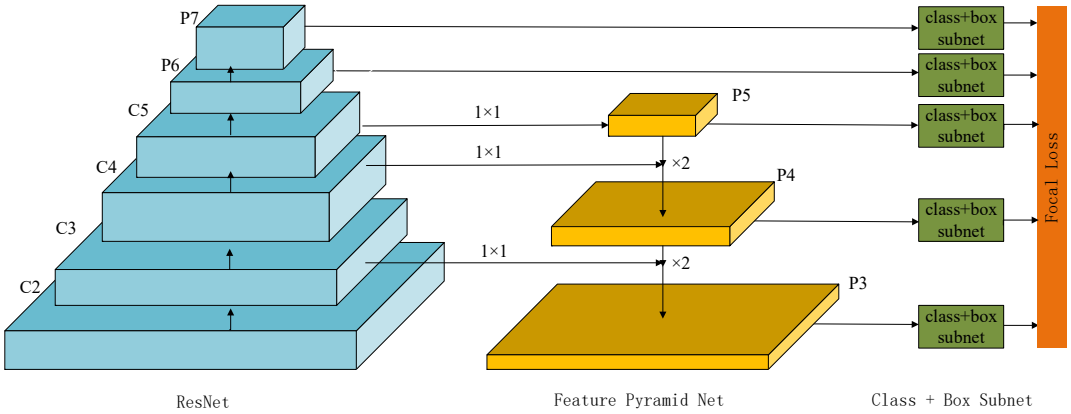

Figure 1. Structure of RetinaNet

The loss function of RetinaNet is the same as other detection algorithms, which can be divided into position loss and classification loss. The calculation of position loss is shown in Formula (1):

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}smooth_{L1}\left(l_i^m-\hat{g}_j^m\right) \tag{1}$$

In which: x_{ij} represents the intersection over union between the i boundary box and the j real box of the k class; l_i^m and $\hat{g}_j^m$ represent four position parameters of the bounding box and four parameters of the real box respectively. The smoothL1 Loss is shown in formula (2):

$$smooth_{L1}=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & others\end{cases} \tag{2}$$

Considering the imbalance between positive and negative samples, RetinaNet adopted Focal Loss as the classification loss. Focal Loss is improved in cross entropy loss function (CE Loss), which is shown in formula (3):

$$CE(y)=-\frac{1}{n}\sum_{t=1}^{n}\left[I\{y=y_t'\}\log\left(p_t\right)\right] \tag{3}$$

In which: n represents the total number of bounding boxes, and y_t' is the correct category corresponding to the t bounding box; p_t is the prediction category of the t bounding box; I is a symbolic function; The judgment conditions are in braces. Focal Loss adds weight before CELoss, as shown in formula (4):

$$FL(y)=\alpha\left(1-p_t\right)^{\gamma}CE(y) \tag{4}$$

In the formula: α and γ are the two weight factors of the loss function. In general, it can be seen from the formula that when the proportion of negative samples is high and the samples are unbalanced, the loss will be greatly reduced under the mediation of the weight factors; However, when the proportion of positive and negative samples is uniform, the loss reduction is small, thus reducing the accuracy reduction caused by the large proportion of negative samples.

3. Improvement of RetinaNet model

In this paper, the problem of small target detection in aerial images is taken as the main research object, based on VISDrone2019 data set, and then the actual situation of small target detection in aerial images is analyzed, so as to adjust the hierarchical structure of the network. The structure of ResNet is improved by using the channel volume integral group with multiple receptive fields. For

different tasks, it can learn the allocation strategy of receptive fields in each layer. The improvement work consists of the following three parts:

3.1. ResNet152

The ResNet network in the RetinaNet model was increased from 101 to 152 layers. Its structure is shown in figure 3. Residual connection network no longer have symmetry, so as to enhance the capacity, the characterization of network to enhance the generalization ability of the model.

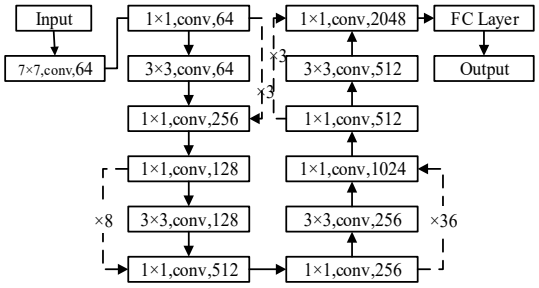

Figure 2. Structure of RetinaNet152

3.2. ScaleNet

In general, when using deep neural network model training, the amount of time is longer, and the model as a whole is bigger,This paper has learned from the thinking of the ScaleNet [17] model, and changed it into the ResNet152 network structure, so as to accelerate and compress the RetinaNet model. ScaleNet uses channel convolution grouping with multiple receptive fields to improve the structure of ResNet. For different tasks, it can learn the receptive field allocation strategy of each layer. In this paper, the 3 × 3 convolution of ResNet152 is replaced by Scale

Aggregation blocks to form a scale network. SA explicitly downsamples the input feature mapping of a group of factors to a small size, and then convolution independently, resulting in different proportions. Finally, the SA block samples the multi-scale representation up to the same resolution as the input feature map and connects them in the channel dimension. Due to the downsampling in each SA block, the sampling density in the spatial domain is effectively reduced, so that the multi-scale architecture has higher computational efficiency and can capture a larger scale (or receptive field) range, as shown in figure 3.

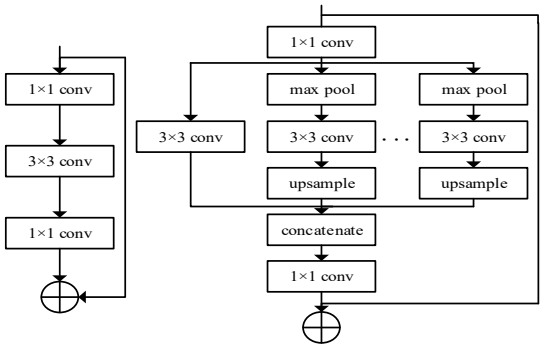

Figure 3. Structure of ScaleNet

3.3. Feature fusion

In many works of deep learning (such as target detection and image segmentation), fusion of features of different scales is an important means to improve performance. In order to make the RetinaNet model more suitable for small target detection in aerial images, it is further improved in this paper. Through the observation of the data set used, the image contains a large number of targets. With the

deepening of the feature extraction network, the pixels contained in the small target object will be reduced to ten or even single digits, which is not conducive to its detection and recognition, so the P2 feature layer is added to the original RetinaNet model to increase the range of the network receptive field through the fusion of stratum features, so as to improve the ability of recognition and detection of smaller objects. The specific structure is shown in figure 4.

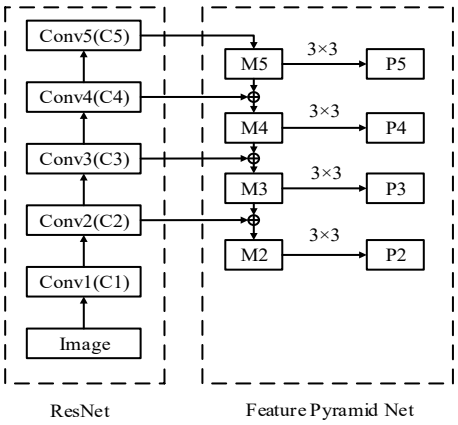

Figure 4. Improved RetinaNet model structure

4. Simulation comparison and analysis

4.1. Experimental prcess

The VISDrone [18] data set was collected by the AISKYEYE team of Tianjin University. In the VISDrone2019 dataset used in this paper, the images are all captured by uav, which is very consistent with the features of aerial images required by the algorithm in this paper. Through the statistics of the target scale, it can be known that the target scale in the data set distributed within 32 accounts for a large proportion by taking the square root of the area as the measure. Small targets refer to objects whose area accounts for one thousandth of the original image, or whose pixel is less than 1024. Therefore, detection of small targets in the data set has a great impact on the final result. In the detection, the RetinaNet superparameters used in this paper include the following four: The ratio of length to width is set to {0.5, 1, 1, 2, 3}. The ratio of scale is set to {2-1/2,20, 21/2}. In order to improve the matching rate of the scale prediction frame to the small target, the positive and negative sample thresholds are set to 0.4 and 0.3 respectively. Although this setting improves the matching rate of the scale prediction frame, it increases the imbalance of the positive and negative sample ratio. Therefore, it is necessary to change the two weighting factors of Focal Loss to improve this situation. Therefore, the Focal Loss parameters are set as $\alpha = 0.25$ and $\gamma = 3.0$. In order to verify the detection performance of the improved network model, the following four models were tested under the same experimental environment. The training weights of the improved RetinaNet on the COCO [19] target detection dataset are used as the pre-training parameters of the model. In order to make the experimental results more in line with the needs of the actual situation, random horizontal flip and random rotation operations are applied to the image, and then the image is standardized on the data set to make it easier to obtain the generalization effect after training. In the experiment, the initial learning rate of each network model training is 0.02, the weight attenuation coefficient is 0.0001, and the batch_ size value is set to 16. AP represents the average over all IOU thresholds and categories, AR represents the average over categories and IOU and categories for a fixed number of maximum recalls. In order to verify the improved model detection effect and detection speed, statistics of each model in the experiment and the average value generated a prediction picture. The time required to verify and compare the detection accuracy and speed of each model, the results are shown in Table 1.

Table 1. Detection effects of different models

Network Model	AP/%	Time/ms
RetinaNet	18.98	85
＋ResNet152	20.21	89
＋ScaleNet	23.17	80
＋P2	25.19	82

It can be seen from Table 1 that with the step-by-step improvement of the network model, the performance of the algorithm has increased significantly. Compared with the original RetinaNet, the improved RetinaNet algorithm has an AP increase of 6.21%, indicating that the feature fusion module is beneficial to extract more effective target features. The detection time is also in practice. Figure 6 shows the detection results of RetinaNet and the improved RetinaNet. Comparing the two figures, it can be seen that the improved RetinaNet model detects more objects than the original RetinaNet model, and has better performance for detecting small target objects with multi-scale changes. In line with the expected effect of the model.

(a) (b)

Figure 5. Comparison of RetinaNet and improved RetinaNet .(a)Test result of RetinaNet; (b)est result of improved RetinaNet

Aerial images have many characteristics, there are certain light and dark changes, there are scenes such as overexposure and insufficient light. The density of objects in the scene changes greatly, there are scenes with extremely high target density, the image field of view is large, the single target is small, the background is extremely complex, and the phenomenon of target occlusion is also common in the data set. Figure 6 shows the visualized image of the detection results of the improved RetinaNet in various scenarios.

Figure 6. Improved RetinaNet detection results in a variety of scenarios

4.2. Analysis of experimental results

Table 2 shows the detection effects of various algorithms on targets of different scales. It can be seen that the improved RetinaNet's detection accuracy for small targets significantly exceeds the original RetinaNet algorithm, and the target detection performance of various scales has been greatly improved. After adjusting only the scale prediction frame without making other changes, the RetinaNet algorithm is named A-RetinaNet as the intermediate state of the algorithm comparison. It can be seen from the table that the algorithm has a certain improvement in the detection effect of small targets after adjusting the scale prediction frame. The set scale prediction frame parameters and Focal Loss weight factors are suitable for the distribution characteristics of the aerial image targets in the VISDrone dataset. The improved RetinaNet model is similar to the one-stage detection algorithm in detection speed, and the detection accuracy is also higher, which is more suitable for target detection in aerial images.

Table 2. Comparison of detection effects of targets of different scales

Method	APsmall/%	APmedium /%	APlarge/ %	ARsmall/ %	ARmediu m/%	ARlarge /%	F1-score
RetinaNet	7.31	23.89	36.75	10.29	26.58	42.35	32.66
A-RetinaNet	9.68	25.07	38.10	14.46	31.25	44.69	37.22
Improved RetinaNet	12.20	34.26	54.63	16.88	47.83	59.22	51.13

5. Conclusion

Through the analysis and comparison of common one-stage and two-stage target detection algorithms, it is found that their detection accuracy is not high in aerial image target detection. RetinaNet is selected as the benchmark model and improved to adapt to aerial image target detection. Through statistical analysis of VISDrone2019 aerial image data set, it is found that small target detection is the key to improve the accuracy of aerial image target detection. On this basis, the scale prediction box is adjusted to make it more suitable for aerial image targets. The algorithm uses ResNet152 to extract high-quality detection features, introduces ScaleNet to accelerate and compress the model, and adds P2 network. The range of receptive field of the whole network is increased. The experimental results on the data set show that the average accuracy value generated during target detection is improved by 6.21% before comparison and improvement. Compared with the original algorithm, the improved algorithm has significantly enhanced various indexes, and the detection effect on aerial images of different target scenes in the data set meets the expected expectations. The improved algorithm needs to improve the detection effect of sample categories which account for a small proportion of the whole, and there is still room for improvement in the detection accuracy of small-scale targets. Under the condition of meeting the performance requirements of the algorithm, simplifying the network, using super-resolution reconstruction and Anchor Free, and designing a more effective feature extraction network to further improve the detection performance are the main objectives of the next research.

References

[1] Zhang H, Wang K F, Wang F Y. Advancesand perspectives on applications of deep learning in visual object detection. [J].Acta Automatica Sinica, 2017, 43(08):1289−1305.

[2] ThYuan C, Liu Z X, Zhang Y M. UAV-based forest firedetection and tracking using image processing techniques [C] // 2015 International Conference on Unmanned Aircraft Systems (ICUAS), June 9-12,2015, Denver, CO, USA. New York: IEEE,2018:639-643

[3] Sun H M, Cao T G, Dai Z X, Peng P. Research on power capacity monitoring technology based on dual-channel images [J]. Laser and Infrared, 2019,49(11):1338-1343.

[4] Girshick R, Donahue [J], Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014: 8-13.

[5] Jiang W T, Zhang C, Zhang S X, Liu W J. Target detection based on multi-scale feature map fusion [J]. journal of image and graphics, 2019,24(11):1918-1931.

[6] Chen T M, Fu G Y, Li S Y, Li Y. typical target detection of infrared terminal guidance based on YOLO v3 [J]. advances in laser and optoelectronics, 2019,56(16):155-162.
[7] Girshick R. Fast R-CNN[C]//2015 IEEE Intwrnational Conference on Computer Vision (ICCV), December7-13, 2015, Santiago, Chile. NewYork: IEEE, 2015:1440-1448.
[8] He K M, Zhang X Y, Ren S Q, Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.[J]. IEEE transactions on pattern analysis and machine intelligence,2015,37(9): 1904-1916.
[9] Liu, Zhao Y, Xu M, Chang L. Dynamic partitioned multi-car elevator scheduling based on Fast R-CNN [J]. Control engineering,2019,26(02):208-214.
[10] Ren S Q, He K M, Girshick R,Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.[J]. IEEE transactions on pattern analysis and machine intelligence,2017,39(6)1137-1149.
[11] Si J X, Lin J J, Jiang F, Shen R M . Hand-raising gesture detection in real classrooms using improved R-FCN[J]. Neurocomputing,2019,359:69-76.
[12] Cao Y J, Xu G M, Shi G C. Low altitude armored target detection based on rotation invariant Faster R-CNN [J]. Advances in Laser and Optoelectronics, 2018,55(10):225-231.
[13] Deng Z P, Sun H, Lei L, Zhou S L, Zou H X. High-resolution Remote sensing image target Detection based on multi-scale Deformation feature convolution Network [J]. Journal of Surveying and Mapping, 2016,47(09):1216-1227.
[14] Xie X L, Li C X, Yang X G, Xi J X, Chen T. Aerial image target detection algorithm based on dynamic sensing field [J]. Acta optica sinica,2020,40(04):107-119.
[15] Xie S, Girshick R, Dollár P, et al. Aggregated resdual for deep neural networks[C] .P roceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.
[16] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networs for object detection[C]. Proceedings of the IEEE conferece on computer vision and pattern recognition. 2017: 2117-2125.
[17] Li Y, Kuang Z, Chen Y, et al. Data-driven neuron alocation for cale aggregation networks[C]. Proceedings of the IEEE conferece on computer vision and pattern recognition. 2019: 11526-11534.
[18] Zhu P, Wen L, et al. Vision meets drones: a challenge [J/OL]. (2018-04-23) [2019-08-28]. https://arxiv.xilesu.top/abs/1804.07437.
[19] Gao Q, Lian Q. Research on insulator target detection in aerial images [J]. Electric measurement and instrumentation,2019,56(05):119-123.