An improved YOLO v3 algorithm for remote Sensing image target detection

Fan Yang1, 2, a, *
1The Engineering & Technical College of Chengdu University of Technology, Leshan, Sichuan Province, 614000, China.
2Southwestern Institute of Physics, Chengdu, Sichuan Province, 610225, China.
aEmail: 17761033182@163.com.
*Corresponding author’s e-mail:17761033182@163.com.

Abstract. With more and more in-depth research on deep learning algorithms in recent years, how to use deep learning method to detect remote sensing images is the key to improving the utilization efficiency of remote sensing data and realizing the transformation from data to knowledge. In this paper, an improved YOLO V3 algorithm is proposed to solve the problems of missed detection and false detection of the original YOLOv3 algorithm in remote sensing image target detection with different size and wide disparity in length and width ratio. first of all, K-means algorithm is used for clustering analysis of data set to obtain the position of anchor box; Secondly, the dilated convolution with expansion rate of 2 is used to replace the general convolution in the feature extraction part; Then four scales are used for prediction; Finally, the improved algorithm is applied to the recognition of bridges, harbors and airports. The results show that the detection performance of the algorithm is improved by about 2% compared with the original algorithm.

1. Introduction
With the improvement of remote sensing technology in recent years, the resolution of remote sensing images is getting higher and higher, and the image contains more abundant information, the background of the target is also more complex. In addition, some uncontrollable natural factors such as atmospheric refraction and curvature of the earth cause unusual shape changes in the image target which undoubtedly increased the difficulty of the remote sensing image target detection. Therefore, it is of great practical significance to improve the target detection rate of remote sensing images by optimizing the target detection model [1].

At present, there are two kinds of target detection algorithms based on deep learning, they are candidate region-based method and regression based method respectively. The former is represented by R-CNN, Fast R-CNN [2], Faster R-CNN [3], R-FCN [4], etc. This method has high detection accuracy but the detection speed is slow, and cannot meet the demand of real-time. The latter is represented by YOLO [5] and SSD [6]. This kind of algorithm has higher detection speed but lower accuracy.

2. YOLOV3
YOLO V3 is improved on the basis of YOLO V2, which can be summarized in three aspects. the network structure is changed from darknet-19 to Darknet-53; Using multi-scale features to detect objects;
The object classification method uses Logistic to replace Softmax, which can support multi-label object detection.

2.1 Network architecture Darknet-53
Darknet-53 draws on the ideas of ResNet and adds a residual module to the network to help solve the gradient problem of deep networks. Each residual unit (RES_unit) consists of two convolutional units (DBL) and shortcut connections (Figure 2). In the whole network structure, there is no pooling layer and full connection layer. The network is down-sampled by convolution of step 2, and the size of the image will be reduced to half after passing this convolution unit. Each convolution unit consists of a convolution layer, a batch normalization layer, and a Leaky ReLU activation function (Figure 1). The new structure can reduce the loss of target information and improve the detection ability of small targets [7].

2.2 Activation Function
In general, the initial input of data stimulates neurons linearly, but in fact, the expression ability of linear models is far from meeting the requirements. The greatest significance of activation function is to solve nonlinear problems and improve the expression ability of deep convolutional network models. The original algorithm YOLO v3 uses the Leaky ReLU activation function, and its formula is:

\[
f(x) = \begin{cases} 
  x, & x \geq 0 \\
  \alpha x, & x < 0 
\end{cases}
\]  

(1)

3. Improved YOLO V3 algorithm
3.1. Select the anchor box by clustering
Due to the big difference in the data set, the size of the anchor box obtained by the original clustering of YOLOv3 is not suitable for the remote sensing image of this article, so the data is re-clustered. The experiment shows that as shown in Figure 3, as the number of anchor boxes increases, the Intersection-over-Union increases, at the same time, the complexity caused by the increase in the number of anchor boxes is also increasing. Finally, weigh the number of IOU and the number of anchor boxes, choose the number of the boxes is 12, and the 12 anchor boxes are (16,16), (29,27), (27,126), (38,48), (62, 36), (66,76), (86,140), (120,61), (163,268) (206,130), (330,253), (380,134). On each scale, each cell predicts three bounding boxes with the aid of 3 anchor boxes.
3.2 Using Dilated convolution

The original network structure Darknet-53 uses a large number of 1*1 convolution kernels and 3*3 small convolution operations instead of pooling operations. Although this reduces the amount of parameters, the resolution of the feature maps in the first few layers is higher. In the case of small size, the local receptive field is too small to capture good features, which affects the model's detection rate of small targets. If it is replaced with a larger convolution kernel, it will increase the amount of parameters and calculations. Therefore, in order to reduce the amount of parameters while maintaining accuracy, the general convolution is replaced by dilated convolution with an expansion rate of 2. The network structure is shown in Figure 4.

| Type      | Filters | size  | output   |
|-----------|---------|-------|----------|
| Conv      | 32      | 3x3   | 512x512  |
| Conv      | 64      | 3x3   | 256x256  |
| Conv      | 128     | 3x3   | 128x128  |
| Dilated Conv | 64      | 1x1   | 256x256  |
| Conv      | 256     | 3x3   | 64x64    |
| Dilated Conv | 256     | 1x1   | 64x64    |
| Conv      | 512     | 3x3   | 32x32    |
| Dilated Conv | 512     | 1x1   | 32x32    |
| Conv      | 1024    | 3x3   | 16x16    |
| Dilated Conv | 1024    | 1x1   | 16x16    |
| Conv      | 2048    | 3x3   | 8x8      |
| Dilated Conv | 2048    | 1x1   | 8x8      |

Figure 4. The improved network structure

Figure 5. The improved convolutional components

Firstly, the first convolutional layer uses 32 convolution kernels (filters) with a size of 3*3 to filter the input image with a resolution of 512x512. Next, take the output of the previous convolutional layer
as input. Using Sixty-four convolution kernels with a size of 3*3 and two steps to filter them to achieve down-sampling, adding residual blocks like the original YOLOv3 algorithm to increase the depth of the network. A dilated convolution layer with a rate of 2 and a size of 3*3 is used to replace the original general convolution in the larger feature map in the above. It does not increase any computation and parameter number, nor does it change the size of the output feature graph of any layer. Afterwards the size of the obtained feature graph is 256*256. Then five groups of 2*, 8*, 8*, 4*, 4* networks containing residual blocks were executed to obtain feature maps with resolution of 128*128, 64*64, 32*32, 16*16 and 8*8 respectively. Each residual block is similar except that the number of convolution kernels and the size of the feature graph are different in the network composed of five groups of residual blocks. Meanwhile, down-sampling is performed in the convolution layer above each rectangular box with a step size of two pixels. Furthermore, a batch standardization layer (BN Layer) is added to all convolutional layers of the network to help standardize the network (as shown in Figure5). In the end, the 64*64, 32*32, 16*16, 8*8 feature maps are fused with the up-sampling feature maps to form a feature gold pagoda for remote sensing image prediction.

3.3 Increased scale prediction
The original YOLO v3 algorithm defaults to three scale predictions. Since the shallow large-scale feature map of the convolutional neural network can better retain the information of small targets, and the deeper feature map has better target recognition effect, the use of multi-scale prediction can combine the advantages of the two to obtain a complete and clear target, so the improved model in this paper uses four scale predictions (as shown in Figure 6), which are 64*64, 32*32, 16*16 and 8*8. In the feature extraction process, the low-level feature map with high resolution and the high-level map with higher semantics are combined, and a more accurate anchor box is assigned to the small target on the larger feature map.

![Figure 6. The Multi-dimensional prediction structure diagram](image)

3.4 Changing the activation function
Considering that ReLU [8] is more non-linear than Leaky ReLU, which can more effectively alleviate the problem of gradient disappearance and reduce the training time of the model, ReLU is used to replace Leaky ReLU as the activation function. The function formula is:

\[
f(x) = \begin{cases} 
  x, & x \geq 0 \\
  0, & x < 0 
\end{cases}
\] (2)

4. Experimental results and analysis
The hardware and software platforms of the experiment in this article are as follows, operating system: Windows7; CPU: Intel(R) Core (TM) i7-7800 CPU @ 3.50GHz *12 processor; GPU: NVIDIA GeForce
RTX 2080Ti; memory: 16G; programming language: Python; deep learning framework: Tensorflow. The main source of the images used in the experiment is the first season of the "AIIA" Cup (Aerospace Science and Industry Station)-a small sample satellite image typical target recognition competition data set, which contains a total of 6000 pictures, training set, validation set and test set the ratios are 4:1:1. In order to verify the effect of the model, this experiment selected bridges, harbors, and airports with large disparity in length and width, complex background, and rich target information in remote sensing images as target data.

4.1 Model evaluation index
In the field of target detection, there are many evaluation indexes with different research focuses, such as recall rate, precision rate and accuracy rate, etc. The specific definitions are shown in the following table:

| Prediction       | True       | Relevant  | Non relevant |
|------------------|------------|-----------|--------------|
| Retrieved        | True positives (TP) | False positives (FP) |
| Not Retrieved    | False negatives (FN) | True negatives (TN) |

Recall rate: the ratio of the number of detected targets to all targets in the sample set.

Recall = \( \frac{TP}{TP + FN} \) \hspace{1cm} (3)

Precision: The ratio of correctly detected targets to the number of detected targets during target detection.

Precision = \( \frac{TP}{TP + FP} \) \hspace{1cm} (4)

Accuracy: The sum of correctly predicted positive and negative cases divided by the total.

Accuracy = \( \frac{TP + TN}{TP + FN + FP + TN} \) \hspace{1cm} (5)

4.2 Result Analysis
First of all, the collected sample data is preprocessed to obtain a standard sample set that can be directly input at the same time, in order to fully train the network and improve the generalization ability of the model, this article expands the training data set by eight data enhancement techniques: image rotation, cropping, image scaling, adjusting image contrast, saturation, adding salt and pepper noise, Gaussian noise, and increasing exposure rate; secondly, input the sample into the network and calculate it layer by layer, Obtain the error from the loss function, perform back propagation, and adjust the parameters of each layer of the network by using the stochastic gradient descent method; Then, a model file containing weighted parameters will be saved, after the network is trained. the trained network model will be applied to the remote sensing image detection. Execute the detection command to get the name and comprehensive score of each target. Finally, from 1000 randomly selected test images, the accuracy and recall rate of the improved model and the original YOLOv3 were compared to evaluate the detection effect of the model.

In this experiment, three types of bridges, airports and harbor in remote sensing images were used as typical small target detection objects. Table2 and Figure7 show the comparison of detection before and after improvement.
Table 2. Comparison between the original algorithm and the improved algorithm

| Target species | The evaluation index | the original YOLOv3 | the improved model |
|----------------|----------------------|--------------------|-------------------|
| Bridge         | Recall               | 80.21%             | 82.37%            |
|                | Precision            | 70.32%             | 72.91%            |
| Airport        | Recall               | 81.15%             | 83.87%            |
|                | Precision            | 74.30%             | 76.02%            |
| Harbor         | Recall               | 78.44%             | 79.82%            |
|                | Precision            | 82.51%             | 83.97%            |

It can be seen from the table that the improved algorithm has a small improvement in the detection effect of seaports, but a significant improvement in the detection effect of bridges, airports and other targets with very small sizes. By comparing Figure 7 (a) and Figure 7 (b), it can be found that the detection rate of the original YOLO V3 algorithm in the test set is lower than 80% before the number of iterations reaches 80,000, the accuracy rate is about 80% on average after 120,000 iterations but most of them are lower than 80%. However, the accuracy of the improved algorithm model has reached 80% after about 70,000 iterations. Moreover, the accuracy is generally higher than 80% after 120,000 iterations, even reaching 82-84%, which is more than 2% higher than the original YOLO V3 algorithm.

Several groups of pictures were selected to intuitively compare the differences before and after improvement.
Figure 8. Comparison of missed detection results

(a) before improvement (b) After improvement

Figure 9. Comparison of false check results

(a) before improvement (b) After improvement

Figure 10. Comparison of Check box results

(a) before improvement (b) After improvement
By comparing the detection results of YOLOv3 and the improved model in this article, it can be found from Figure 8 that there are two bridge targets in the original image. Figure 8(a) The original YOLO model misses a target and the target synthesis is only 0.41. Figure 8(b) detects the two targets of the original image, and the comprehensive scores are 0.82 and 0.39 respectively. It can be clearly seen that the model constructed in this paper alleviates the problem of missed detection, and the comprehensive score has also been improved to a certain extent; Figure 8 There is only one target airport in the original picture, but there are two targets in Figure 9(a): airport:0.90; bridge:0.33, Figure 9(b) correctly detects only one target is airport:0.54. The original method can be found In the detection result, other targets are mistakenly detected as bridges; Figure 10(a) shows the situation where the anchor point frame crosses, and Figure10(b) shows After the size of the cluster anchor frame This problem is effectively alleviated; the number of targets detected in Figure 11(a) is two, and scores are 0.36 and 0.47 ; the number of targets detected in Figure 11(b) is 5 with scores of 0.40, 0.67, 0.63, 0.40, 0.37; Figure 11(c) shows that the number of detected targets is 5, and the target scores are 0.70, 0.77, 0.66, 0.69, and 0.81. Figure 11(d) also detected the number of 5 targets, and the scores are 0.85, 0.74, 0.82, 0.75, 0.99. It can be seen that the detection effect of Figure 11 (b) and Figure 11 (d) is significantly better than Figure 11 (a) and Figure 11 (c). Through the comparison results of the above detection results, it can be found that the model constructed in this article has missed detection (Figure 8), false detection (Figure 9), and inaccurate detection frame positioning (Figure 10) and low detection rate of small targets in the original model. The problems of (Figure 11) have been improved to a certain extent.
5. Conclusion
This paper has proposed an improved remote sensing image detection algorithm based on YOLO v3. First, the anchor boxes suitable for remote sensing data by using the K-Means algorithm to clustering have been obtained, it can return to the real target faster and more accurately; Secondly, Taking into account the inadequacy of the original algorithm for small target detection, the dilated convolution was introduced in the feature extraction layer of Darknet53, which expanded the local receptive field of the convolution kernel without increasing the amount of parameters. Then the original three-scale detection was increased to four scales; finally, test results on three types of target data of bridges, harbors and airports show that the improved algorithm has improved the detection accuracy, the missed detection rate is controlled and a good convergence effect is also ensured.

Acknowledgments
This article is one of the phased achievements of the Chongqing Social Science Planning Youth Project "Measurement of Chongqing's Industrial Ecological Level and Improvement Strategies from a Global Perspective" (project number:2020QNJJ24)

References
[1] Chen L., Zhang S.J., Wang Y.X., Remote Sensing Image Detection based on improved YOLOv3 [J]. Journal of Small Microcomputer Systems, 2020, V. 41(11):83-86.
[2] Girshick, R. (2015) Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 1440-1448.
[3] Ren, S., He K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6):1137-1149.
[4] Dai, J., Li, Y., He, K., Sun, J. (2016) R-FCN: Object Detection via Region-based Fully Convolutional Networks. Advances in Neural Information Processing Systems.
[5] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. Computer Vision & Pattern Recognition. IEEE.
[6] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., (2016) SSD: Single Shot MultiBox Detector. European Conference on Computer Vision.
[7] Redmon, J., Farhadi, A., (2018) YOLOv3: An Incremental Improvement. arXiv e-prints.
[8] Zeiler, M.D., Fergus R., (2013) Visualizing and Understanding Convolutional Neural Networks. European Conference on Computer Vision. Springer International Publishing.