Application of Improved YOLO V3 Algorithm for Target Detection in Echo Image of Sonar under Reverb

Fang Jie, Wang Pingbo
(School of Electronic Engineering, Naval Engineering University, Wuhan, Hubei 430000)

Abstract: aiming at the problems of low detection rate and high false alarm rate of small and medium targets in images, a YOLO V3 improvement method is proposed and applied to the detection of small targets. Because the small target occupies less pixels and the feature is not obvious, it is proposed to up-sample the 8 times downsampling feature map of the original network output, splice the 2 times upsampling feature map with the second residual block output feature map, and establish the feature fusion target detection layer with 4 times downsampling output. In order to gain More small target feature information is added to the second residual block DarkNe-t53 the YOLO V3 network structure. k-means clustering algorithm is used to cluster the number of target candidate boxes and the aspect ratio dimension. The results show that the improved YOLO V3 algorithm can effectively detect small targets and improve the recall rate and detection rate of small targets.

1. Introduction
At present, the target detection\textsuperscript{12} algorithm based on deep learning has been widely used in various neighborhoods such as military and civilian\textsuperscript{1}. Among them, YOLO V3\textsuperscript{4} is a mature algorithm with high detection rate and accuracy. According to the practical application requirements of different fields, it is necessary to improve the algorithm specifically in order to achieve the ideal detection effect in practice. In this paper, it is used to detect the target in sonar echo image, which is different from the conventional sonar target detection. Different types of uneven objects contained in the ocean form discontinuities in the physical properties of the medium, and the part of the sound energy irradiated to it is radiated back to form the scattering of sound waves. The sum of these scattering components from all scatterers is called reverberation. It is reflected in the sonar echo image that there are some random aggregation points or bar interference. In view of the fact that the real target also shows a highly similar shape in the sonar echo image, It is difficult to achieve better detection effect in the traditional signal processing algorithm and target detection method. The application of deep learning aims to effectively extract the inner and deep features of the target through neural network to realize the distinction between the target and reverberation interference, and to open up a new feasible direction for the development of this field.

To improve the recall rate and accuracy of sonar echo target detection in reverberation background, the simulation target is added to the measured ocean reverberation background echo based on the regression-based target detection model YOLO V3. A series of sonar echo images are processed to obtain a batch of sonar echo images as network training data sets. first, the 8-fold down-sampled feature map output from the original network is up-sampled twice, and the 2-fold up-sampled feature map is spliced with the output of the second residual block to fuse the feature semantic information in
the high-level two residual units are added to the second residual block DarkNet-53 the YOLO V3 network structure to obtain accurate position information in the low-level network. Finally, to avoid the gradient vanishing, the 6 convolutions before the target detection output layer of the YOLO V3 network are transformed into 2 convolutions and 2 residual cells. Comparing the improved network structure with the YOLO V3, the results show that the recall rate and detection accuracy of the improved network structure are obviously improved.

2. YOLO V3 network presentations

YOLO V3 was first proposed at the beginning of 2018 with major structural changes, most notably the use of multiple independent logical classifiers instead of softmax functions, and the use of similar FPN methods for multi-scale prediction. YOLO V3 use DarkNet-53 basic network, its model has 106-layer network, the deeper network level makes the more important and popular structure in the current advanced model can appear on the YOLO V3, including hop layer connection and residual module5, multi-dimensional detection, and upsampling and feature fusion process. A schematic diagram of the structure of the YOLO V3, as shown in figure 1:

![Figure 1 YOLO V3 Network structure diagram](image)

YOLO V3, the input images were sampled 5 times, and the target was predicted in the last 3 times. A feature map of 3 scales is included in the last 3 downsampling, and the detection is completed using a convolution kernel of 1×1. Of the three feature maps for detection, two are obtained after upsampling and feature fusion. The small feature map provides deep semantic information, the large feature map provides the location information of the target, and the small feature map is fused with the large feature map after upsampling. Therefore, the model can detect both large-scale and small-scale targets.

3. Improved YOLO V3 detection model

The self-made sonar target echo data set has distinct particularity. The target in the data map is based on the measured ocean background, using the method of underwater target modeling, simulating the target at the specified distance and angle, then processing the signal echo by matching filtering and other signal processing algorithms, and finally drawing the echo image of the sonar. The echo image has the advantages of clear picture and consistent size, but at the same time, there are some shortcomings such as the small proportion of pixels in the whole picture, the high similarity with the background, the lack of texture features and the weak correlation between pixels. Therefore, in order to avoid practical detection, the target appears on the segmentation boundary. In order to avoid practical detection, the same image is segmented into Y images. Finally, the appropriate size image is extracted at the intersection point. The original echo image can be processed by the above three steps to obtain the trained sample and the real target image to be detected. However, the anchor boxes
defined by the original network and the hierarchical structure of the network are not completely suitable for the research object of this paper. Therefore, it is necessary to use the K-means clustering algorithm to cluster the targets in the self-made data set, and then modify the hierarchical structure of the network according to the characteristics of the targets.

3.1 Data processing and anchors of clustering data sets through k-means algorithms

YOLO V3 draws lessons from the idea of anchor boxes used in the Faster-RCNN, the purpose is to count the most common anchor boxes shapes in the training set, which can be roughly divided into three shapes: flat long shape, thin high shape and square-like shape with close width and height ratio. Determination of the initial anchor boxes scale will directly affect the detection accuracy and speed of the network to the target. K-means clustering algorithm is used to cluster the width and height of the target frame in the training sample. Average overlap degree (Avg IOU) is used as a measure of target clustering analysis to cluster the self-made sonar target echo data set. Avg IOU objective function of clustering can be expressed as

\[ f = \arg \max_{\text{kn}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_k} I_{\text{IOU}}(A, B)}{n} \]

The A represents the sample, that is, the target in the ground truth; the B represents the center of the cluster; \( n_k \) the number of samples in the k cluster center; k represents the number of clusters; \( I_{\text{IOU}}(A, B) \) represents the intersection ratio of the cluster center box and cluster box; i represents the sample number; j represents the sample number in the cluster center. Set k=1~12, cluster analysis of data set samples, get the relationship between k and Avg IOU as shown in figure 2. As the k value increases, the objective function changes more and more smoothly, Change inflection point can be considered as the best number of anchor boxes. At a k value greater than 8, The curve starts to smooth, and so the number of anchor boxes chosen is 8. It can accelerate the convergence of loss function, It can also eliminate the error caused by the candidate box. The size of the corresponding prediction box is set to 8 clustering centers, The width and height of the cluster center on the homemade data set are (40, 76), (42, 58), (45, 90), (50, 66), (55, 103), (58, 82), (61, 56), (66, 68).

![Figure 2 k-means Cluster analysis results](image)

3.2 Improved YOLO V3 detection model

YOLO V3 network uses feature maps of 8-fold downsampled output to detect small targets. This means that when the target is less than 8 pixel × 8 pixel, The network's prediction of the target is difficult, The detection ability of the target detection layer with 8 times down sampling is limited. In order for the network to obtain more feature information of small targets, Increasing the detection rate for small targets, Using the 4-fold down-sampling feature map from the original network, Because it
contains more location information for small targets. A 4-fold down-sampling feature map of YOLO V3 output, A 2-fold up-sampled feature map is spliced with a 4-fold down-sampled feature map output from the second residual block in the DarkNet-53, Set up a feature fusion target detection layer with 4 times down sampling output, To detect small targets. Meanwhile, In order to get more low-level small target location information, Add 2 residual units to the second residual block of the original network. The original YOLO V3 algorithm uses three scales to predict voc and coco data sets. Each scale in the coco dataset corresponds to three different size prediction boxes. In order to improve the network recall rate and detection accuracy of small targets, According to the clustering results of sonar echo image training samples, Cancel YOLO V3 output detection on the original 3 scales, The output 4-fold down-sampling feature fusion target detection layer is directly used to detect small targets. The target detection output layer of the YOLO V3 network contains six DBL units and a convolution of 1×1. Inspired by the DSSD network, in order to avoid gradient disappearance and enhance feature reuse, six DBL units are transformed into two DBL units and two ResNet units, as shown in figure 3.

As shown in figure 4, the improved YOLO V3 network structure splices the 4 times downsampling feature map output from the second residual block with the 8 times downsampling feature map after 2 times upsampling. A feature fusion target detection layer with 4 times downsampling output is established to detect small targets. YOLO V3 network will scale or crop the detected image, no matter how large the input image is, it will be scaled or cropped into 416 pixel ×416 pixel image. Since all the data maps have been processed to a uniform size of 494 pixel ×340 pixel in this paper, if the images are scaled or cropped, the resolution of small targets in the original image will become lower or the field of view will become smaller. Directly affect the network detection of small targets. Therefore, the improved YOLO V3 network does not compress or crop the image, maintaining the resolution of 494 pixel ×340 pixel to keep the image resolution of the input network unchanged, thus improving the detection performance of the network for small targets.
4. Experimental Results and Analysis

YOLO V3 target detection at multiple scales is a representative algorithm that can take into account large and small targets and has better detection performance for small targets. Therefore, the proposed improved YOLO V3 algorithm is compared with the YOLO V3 target detection algorithm. Using sonar echo image as data set, the YOLO V3 algorithm and the improved YOLO V3 algorithm are compared. The sonar echo image data set is based on the array level echo image segmentation into 494 pixel × 340 pixel image. The data set contains a variety of highly similar reverberations and submarine noise interference. The average image in the target dataset contains about 3 real targets, and they only account for 2% of the total pixels of the image.

using image data sets with resolution of 416 pixel × 416 pixel, 80% images were randomly selected for training, and the remaining 20% images were used for testing.

The experimental conditions are as follows: operating system wi ndows7, depth learning framework is keras; CPU i5-9400F, memory is 32 GB, GPU and

4.1 Network training

YOLO V3 and improved YOLO V3 are trained separately, The initial learning rate during the training phase was 0.001, The attenuation coefficient is 0.0005, When the number of training iterations is 10000 and 15000, Reduce the learning rate to 0.0001 and 0.00001, Make the loss function converge further. The image in the data set is enhanced and expanded by means of rotating image and increasing contrast. convergence curves of the loss values during the improved YOLO V3 network training are shown in figure 5.
After about 20000 iterations, the parameters tend to be stable, and the final loss value decreases to about 16.8. From the convergence analysis of the parameters, we can see that the training results of the improved YOLO V3 network are ideal.

4.2 Network testing
Using the measured sonar echo image, two networks are tested to calculate the recall rate and the accuracy of the detection. Target recall and detection accuracy can be expressed as

\[
R = \frac{X_{TP}}{X_{TP} + X_{FN}}
\]

\[
P = \frac{X_{TP}}{X_{TP} + X_{FP}}
\]

In the formula: \(X_{TP}\) the number of correctly detected targets; \(X_{FN}\) the number of undetected targets; \(X_{FP}\) the number of incorrectly detected targets. There are 156 targets in 180 images tested. Two target detection algorithms are used to test on sonar echo dataset. The results are calculated \(R\) and \(P\) calculated respectively. The results are shown in Table 1.

| Detection algorithm | \(X_{TP}\) | \(X_{FP}\) | \(X_{FN}\) | \(R\) (%) | \(P\) (%) |
|---------------------|------------|------------|------------|-----------|-----------|
| YOLO V3             | 129        | 34         | 27         | 82.6      | 79.1      |
| Improved YOLO V3    | 136        | 23         | 20         | 87.1      | 85.5      |

The detection accuracy of the improved YOLO V3 algorithm for small targets is increased from 79.1% to 85.5%, and the recall rate from 82.6% to 87.1% compared with YOLO V3.
Results: figure 6 shows that the original YOLO V3 has missed and wrong detection of small targets in the scene, and the improved YOLO V3 network can effectively detect small targets in the scene, and can avoid the wrong detection of small targets.

5. Conclusion
A YOLO V3 improvement method is proposed and used in small target detection. First, the samples of the data set are clustered and the corresponding clustering centers on the data set are obtained. And then up-sampling the 8 times downsampling feature map of the YOLO V3 output, stitching the 2x upsampling feature map with the 4x downsampling feature map output from the 2nd residual block in the DarkNet-53, A feature fusion target detection layer with output of 4 times down sampling is established. Finally, In order to get more information about small target features, Two residual cells are added to the second residual block DarkNet-53 the YOLO V3 network structure, The six DBL units in front of the target detection output layer of the YOLO V3 network are transformed into two DBL units and two ResNet units. Experimental results show that, An improved YOLO V3 algorithm improves the recall rate of small targets and the average accuracy of detection. But, uh, And the improved YOLO V3 algorithm is still far from the real-time distance engineering application. How to reduce network structure and computational cost without reducing detection performance, will be the main research direction in the future.
References

[1] Lionetto L, Casolla B, Mastropietri F, et al. Application research of 3D imaging sonar system in salvage process [J]. Applied Mechanics and Materials, 2014, 643(8): 279-282.

[2] Marszal J, Salamon R. Detection range of intercept sonar for CWFMsignals [J]. Archives of Acoustics, 2014, 39(2): 215-230.

[3] Chuang M C, Hwang J N, Ye J H, et al. Underwater fish tracking for moving cameras based on deformable multiple kernels [J]. IEEE Trans on Systems, Man, and Cybernetics: Systems, 2017, 47(9): 2467-2477.

[4] Karoui I, Quidu I, Legris M. Automatic sea-surface obstacle detection and tracking in forward looking sonar image sequences [J]. IEEE Trans on Geoscience and Remote Sensing, 2015, 53(8): 4661-4669.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: unified, real-time object detection [C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2016: 779-788.

[6] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition [C]. Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition Workshops (C- VPRW), Las Vegas, NV, F Jun 26-Jul 01, 2016. IEEE: New York, 2016.

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards realtime object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.