Small Boat Detection for Radar Image Datasets with YOLO V3 Network

Guanqing Li*, Zhiyong Song, Qiang Fu
College of Electronic Science, National University of Defense Technology, Changsha, China
*Email address: (liguanqing09@nudt.edu.cn)

***Abstract*—Small boat detection under the influence of sea clutter is usually difficult, especially for the low-resolution pulse Doppler Radar, as the amplitude of the boat is covered by sea clutter in the time domain, and the spectrum of the boat and sea clutter overlap in the frequency domain. This paper proposes a new method for the detection task based on time-frequency analysis and YOLO V3 network. This method can automatically extract the characteristics of sea clutter and boat in time-frequency images and complete the classification task. The accuracy of the classification is 94.89 percent, which has an improvement of 14.39% than the accuracy of LeNet-5. The measured data verified the method.**

***Keywords*—YOLO V3, Time-Frequency Analysis, Small Boat Detection, Radar, Deep Learning**

I. Introduction

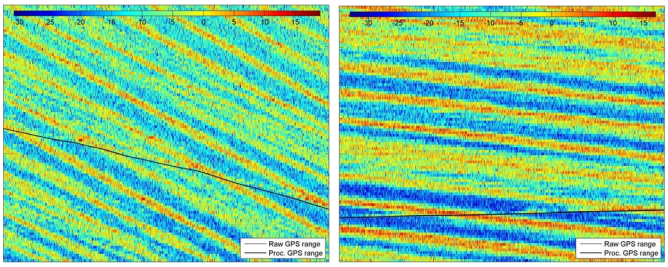
China has more than 3 million square kilometers of sea area and 18,000 kilometers of coastline, ensuring marine safety is an important aspect of maintaining national security. Radar can detect and track the targets without the influence of time and climate. It is an indispensable part of the marine military.

The sea observation radar will encounter severe disturbances of the sea surface radar echo when it is working. This interference signal is generally called sea clutter. Sea clutter is affected by radar parameters, wave height, wind direction, ocean currents, rainfall, and seawater dielectric constant [1]. Sea clutter is much more complicated than ground clutter. In addition, if the target to be observed is small, the target echo signal is often submerged in the strong sea clutter, making the detection problem more difficult.

Su et al. [2] use deep CNN to classify different sea states and polarizations through the IPIX measured data. However, for low-resolution pulse Doppler radar, when the target and sea clutter are mixed in both the time domain and the frequency domain, how to accomplish the small target detection task is still an urgent problem to be solved.

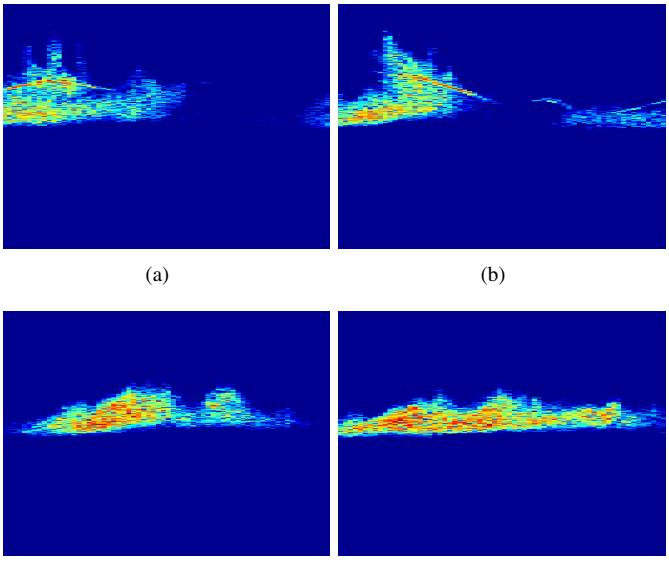
The experimental data is from the publicly available Fynmet dynamic radar data from the South African Scientific and Industrial Research Council (CSIR) in Overberg. The CSIR radar is vertically polarized at a frequency of 9 GHz with a pulse repetition rate of 5 kHz, a range resolution of 15 m and a boat size of 3-5 m.

The image below is from CSIR Fynmet dataset TFC15_009 and TFC15_011. The first 132 seconds of returning data are shown in Figure 1, which represents the amplitude of the radar echo from the periodic waves and targets of the Overberg coast near Cape Town, South Africa.



(a) (b)

Fig. 1. The horizontal axis represents 96 distance ranges, each of which is 15 meters. The vertical axis represents the pulse, a total of 57,334 pulses. The black solid line represents the GPS trajectory of the boat. The target echo is completely annihilated in the strong sea clutter.



(a) (b)

(c) (d)

Fig. 2. Time-frequency image obtained by the short-time Fourier transform for the original echo data. The horizontal axis represents time, 0-13.46s. The vertical axis represents the Doppler shift, which is -800 Hz to 800 Hz from bottom to top. The first line contains the target and the second line does not.

978-1-7281-2345-5/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on October 11,2020 at 03:48:12 UTC from IEEE Xplore. Restrictions apply.

We take a column of the original echo data for the short-time Fourier transform, and some of the resulting time-frequency images are shown in Figure 2. It can be seen from Figure 2(a) and 2(b), sometimes the sea clutter and the Doppler shift of a boat is mixed, and the Doppler frequency shift of sea clutter is very strong.

In this paper, we propose a new method for the detection task based on time-frequency analysis and YOLO V3 network. The measured data shows that our method can effectively complete the classification task. At the end of the article, the result of our experiments is compared with the classical convolutional neural network LeNet-5 [3].

II. Method

A. Overall Architecture

The overall architecture of our method is shown in Figure 3. The raw data is a two-dimensional complex-valued matrix. The horizontal axis data has been processed by pulse compression, representing the distance range, and the vertical axis represents the sequence of the pulse. The resulting images generated by STFT are saved as three-channel images in PNG format by pseudo-color processing.

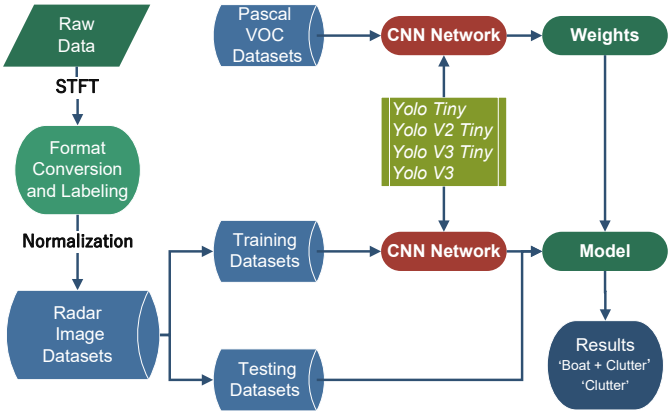


Fig. 3. The architecture of our approach. The left and right part of the picture is the procedure of generating time-frequency image datasets and how to detect small boats through convolutional neural networks, respectively. The upper right part of the figure is a flow chart for transfer learning on the Pascal VOC dataset.

During the forward training process, the images from the training datasets are put into the CNN, including a series of convolutional layers, max-pooling layers, and fully connected layers to produce suggested features. We get the initial weight of the network by training on the Pascal VOC dataset. During the back-propagation process, we calculate the loss. We put the testing image into the saved model to predict whether the image is 'sea clutter' or 'boat + sea clutter'.

B. YOLO V3 Network

The network architecture of YOLO V3 is shown in Figure 4. It is improved on the basis of YOLO [4] and YOLO V2 [5], mainly in two aspects: one is to carry out multi-scale prediction similar to FPN; the other is to use a better basic classification network and classifier similar to ResNet. The figure shows the structure of the YOLO V3 network [6], consisting of 75 Convolutional Layers, 20 Shortcut Layers, and 2 Up sample Layers. The improvement of YOLO V3 is mainly reflected in the following two aspects.

1) Multi-scale prediction: The low-level feature semantic information is relatively small, but the target location information is accurate; the high-level feature semantic information is rich, but the target location information is relatively rough. In YOLO V3, three boxes are predicted for each scale, and the anchor design method still uses clustering to obtain 9 cluster centers, which are divided into three scales according to their size. YOLO v3 adopts Upsample, which combines three scales (13*13, 26*26 and 52*52), and independently performs detection on the fusion feature maps of multiple scales, and finally improves the detection effect of the target.

2) Basic network Darknet-53: Unlike Darknet-19, YOLO V3 uses a 53-layer convolutional network that is superimposed by residual units. There are two main changes for the Darknet-53 convolutional layer. One is to replace the original block with a simplified Residual block. The second is to use concatenated convolution to increase the channel. The basic network comparison of YOLO v2 and v3 is shown in the TABLE I.

TABLE I
THE BASIC NETWORK COMPARISON OF YOLO V2 AND V3.

Darknet 19				Darknet 53				
Type	**Filer**	**Size**	**Out**	**Tis**	**Type**	**Filer**	**Size**	**Out**
Conv	32	3	224		Conv	32	3	256
Maxp		2	112		Conv	64	3	128
Conv	64	3	112		Conv	32	1	
Maxp		2	56	1×	Conv	64	3	
Conv	128	3	56		Resid			128
Conv	64	1	56		Conv	128	3	64
Conv	128	3	56		Conv	64	1	
Maxp		2	28	2×	Conv	128	3	
Conv	256	3	28		Resid			64
Conv	128	1	28		Conv	256	3	32
Conv	256	3	28		Conv	64	1	
Maxp		2	14	8×	Conv	128	3	
Conv	512	3	14		Resid			64
Conv	256	1	14		Conv	256	3	32
Conv	512	3	14		Conv	64	1	
Conv	256	1	14	8×	Conv	128	3	
Conv	512	3	14		Resid			64
Maxp	64	2	7		Conv	256	3	16
Conv	1024	3	7		Conv	64	1	
Conv	512	1	7	4×	Conv	128	3	
Conv	1024	3	7		Resid			8
Conv	512	1	7		AvgP		Glo	
Conv	1024	3	7		FC		1k	
Conv	1000	1	7		SofM			
AvgP		Glo	1k					
SofM								

In the TABLE I, 'Conv', 'Maxp', 'AvgP', 'SofM', 'Out', 'Tis', 'Resid', 'Glo', and 'FC' means convolutional layer, max pooling layer, average pooling layer, soft-max layer, output size, number of network repetitions, residual layer, global, and fully connected layer, respectively.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on October 11,2020 at 03:48:12 UTC from IEEE Xplore. Restrictions apply.

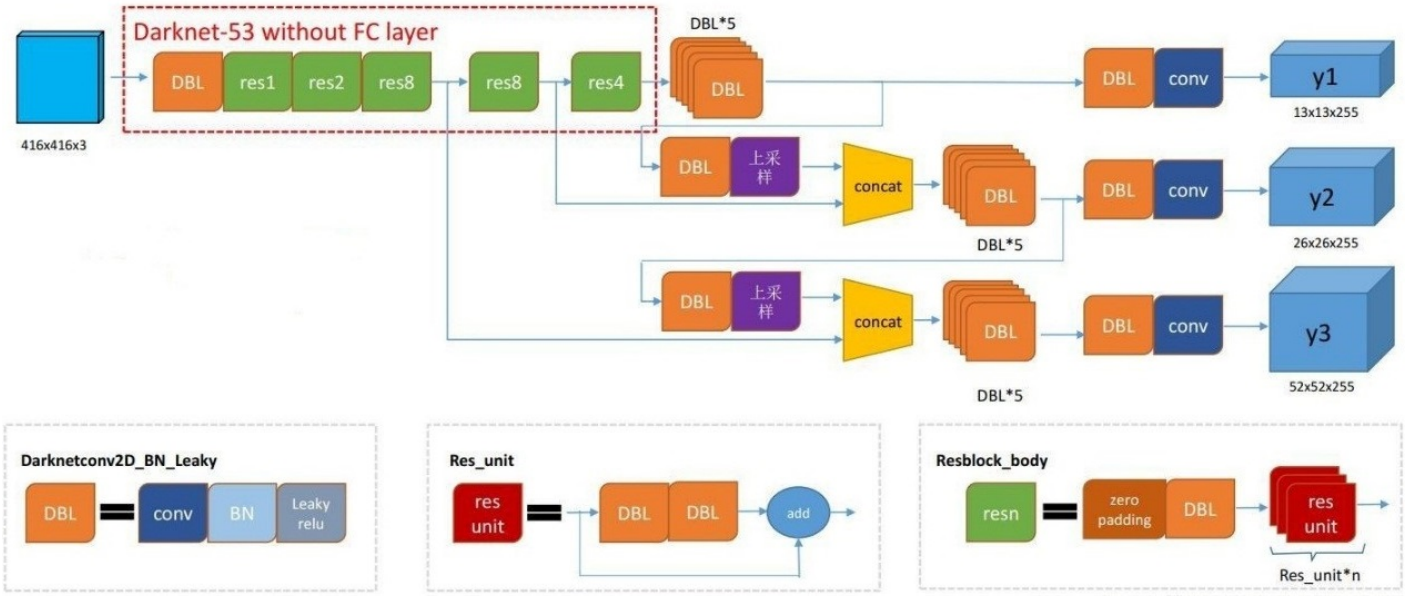


Fig. 4. YOLO V3 network structure.

C. Time Complexity of the Network

The time complexity of the network is analyzed from the size of the input and output of each layer. Only the number of multiplication operations is calculated. The amount of operation can be expressed by the following formula:

$$\mathrm{O}\left(N,F,n\right)=\sum_{\mathrm{i}=1}^{75}N_i^{out}\cdot F_i\cdot\left(n_i\cdot n_i\cdot F_{i-1}\right) \tag{1}$$

Where N_i^{out} is the output image size of the i-th layer network, and F_i is the number of the i-th layer convolution kernel (the number of channels), and n_i is the size of the i-th layer convolution kernel. The total available computation of the YOLO V3 network is 65.290 BFLOPS.

D. Non-maximum suppression algorithm

TABLE II
THE NON-MAXIMUM SUPPRESSION ALGORITHM

NMS Algorithm for Object Detection
Input: $B=\{b_1,b_2,\cdots,b_N\}, S=\{s_1,s_2,\cdots,s_N\}, N_t$
B is the list of initial detection boxes
S contains corresponding detection scores
N_t is the NMS threshold
begin
$D\leftarrow\{\}$
while $B\neq\varnothing$ do
$m\leftarrow$ argmax S
$M\leftarrow b_m$
$D\leftarrow D\cup M; B\leftarrow B-M$
for b_i in B do
If $iou(M,b_i)\geq N_t$ then
$B\leftarrow B-b_i; S\leftarrow S-s_i;$
end
end
end
return D, S
end

Non-maximum suppression (NMS) is widely used in target detection algorithms. Its purpose is to eliminate redundant candidate frames and find the best object detection location. Now suppose there is a set of candidate boxes B and its corresponding scores set S. Firstly, we find the m with the highest score; then we remove the box corresponding to M from B; and then, we add the deleted box to the set D; next, the other boxes in which the box overlap area corresponding to M is larger than the threshold value N_t are deleted from B; finally, the above steps 1-4 are repeated. The pseudocode is in TABLE II.

E. Evaluation index

After the classification, it is necessary to evaluate the classification results. In addition to the commonly used correct rate, the evaluation criteria include recall accuracy, false alarm rate, and missed alarm rate.

Accuracy indicates the proportion of positive samples were correctly classified. The formula is as follows:

$$P_a=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}} \tag{2}$$

Where N_{TP} indicates the number of positive class samples correctly classified, N_{TN} indicates the number of negative class samples correctly classified, N_{FP} indicates that negative class samples are divided into positive classes, and N_{FN} indicates that positive class samples are classified into negative classes.

III. EXPERIMENTS

In this section, we will describe the training process, the testing process, the training parameters after adjusting the parameters, and the experimental results of the test.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on October 11,2020 at 03:48:12 UTC from IEEE Xplore. Restrictions apply.

A. Training

Training the radar dataset with the YOLO V3 network is a step-by-step process:

Step1, A short-time Fourier transform is performed on the original echo data, and the width of the window function is 0.5 s, which is a balance of time resolution and frequency resolution.

Step2, The time-frequency image is marked by the GPS information of the radar data set.

Step3, Let the YOLO V3 network pre-train on Pacasal VOC 2012 to get the network weight file yolov3.weights.

Step4, Load the weights and adjust parameters, including learning rate, momentum attenuation, batch size, epoch.

There are 1212 images in total. 70% of them were used for training and the rest were used for testing. Here, we randomly extract 848 pictures as the training set, the remaining 364 pictures as the test set, and then load the weight file to adjust the parameters. The results of the adjustment are in TABLE III. The results of the adjustment for LeNet-5 and YOLO V2 are also among them.

TABLE III
THE RESULTS OF THE ADJUSTMENT.

Parameter	LeNet-5	YOLO V2	YOLO V3
Learning Rate	1e-3	1e-4	1e-4
Momentum	0.9	0.9	0.9
Batch Size	64	16	8
Epoch	2000	1000	2000
Training Time	30 min	2 h	4 h

We recorded and visualized the loss and accuracy values of YOLO V3 during the 2000 epoch training epochs. Figure 5 and Figure 6 show the curves of loss and accuracy in training, respectively.

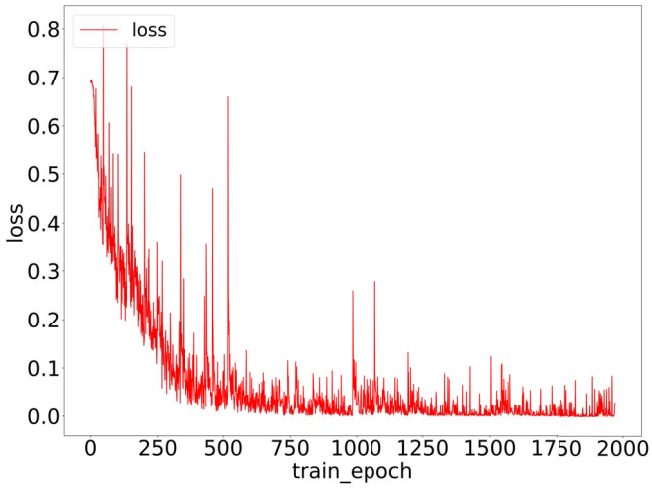


Fig. 5. The curve of loss during training, due to the use of transfer learning, the initial loss value is relatively small.

B. Testing

The following steps are the process of testing the radar time-frequency image.

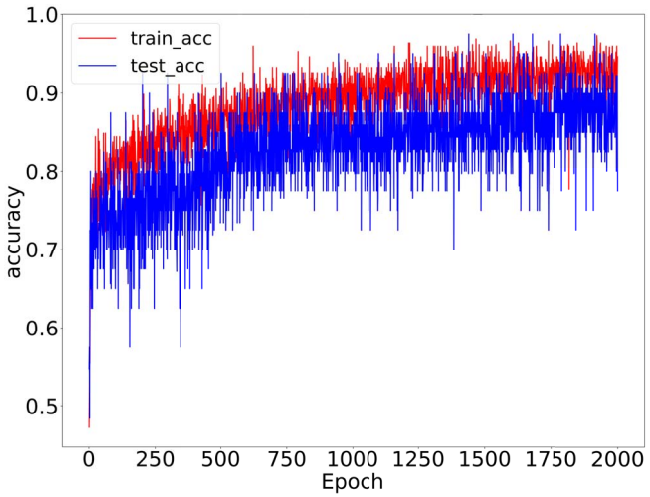


Fig. 6. Accuracy curve during training, red solid line represents training accuracy, and blue solid line represents test accuracy.

Step1, Load the model saved during training, ie the yolov3.pb and yolov3.meta files, and feed the image into the model.

Step2, Use a non-maximum suppression algorithm to discard the extra bounding boxes, leaving at most one bounding box.

Step3, If there is a prediction box, the prediction result is 'sea clutter + boat', otherwise 'sea clutter'.

In step 2, there is one point that needs to be explained. Through the training, each picture can be predicted to get $(13 \times 13 + 26 \times 26 + 52 \times 52) \times 3 = 10647$ bounding boxes, each bounding box contains six parameters: horizontal axis upper left coordinate, vertical axis upper left coordinate, the coordinates of the lower corner of the horizontal axis, the coordinates of the lower right corner of the vertical axis, the confidence of the target in the box, and the probability of the target being a boat. There are too many prediction boxes here, so you need to use the non-maximum suppression algorithm to remove the extra boxes. In addition, in the boat inspection mission, there is only one type of target in the label, the 'boat'. Therefore, we only need to reserve at most one prediction box.

The test system is Linux, the specific version is Ubuntu 16.04, and the CPU processor is a 4-core 3.4GHz Intel i5 7500 processor, which is the same as the training system. The GPU is not used during the test.

A total of 364 pictures of the ship were included in the 182 test pictures, 182 pictures without the ship. The accuracy of recognition is 94.89%, and the evaluation index is shown below:

From TABLE IV we see that the YOLO V3 indicators have the best performance. Compared to LeNet-5, which is directly classified using convolutional neural networks, YOLO V3 has increased by 14.39 and 15.38 percent respectively in accuracy and precision. Its false alarm rate dropped by 6.45 percent. Its false alarm rate dropped by 7.14 percent. And the former is positioned at the same time as the classification, while the

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on October 11,2020 at 03:48:12 UTC from IEEE Xplore. Restrictions apply.

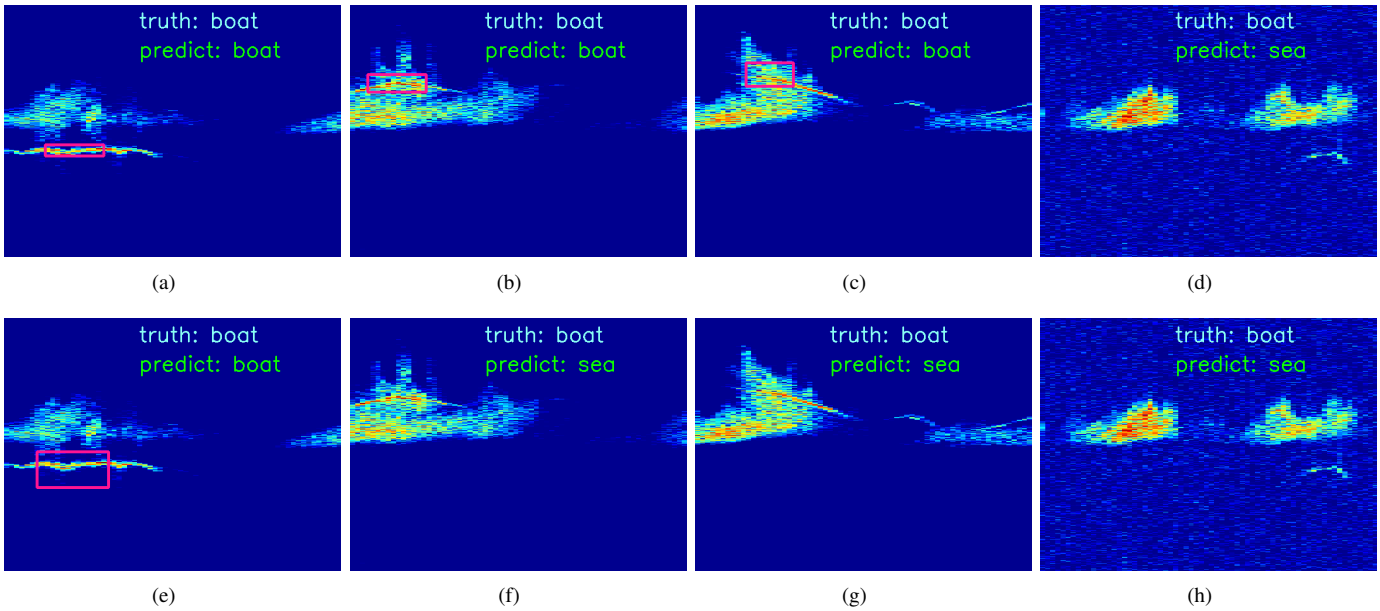


Fig. 7. More detailed results.The first and second lines are the test results of YOLO V3 and V2, respectively.

TABLE IV
THE TESTING RESULT.

Index value	LeNet-5	YOLO V2	YOLO V3
Accuracy P_a	0.8050	0.8267	0.9489
Precision P_p	0.8022	0.8352	0.9560
Missing Alarm P_{ma}	0.1978	0.1648	0.0440
Recall P_r	0.8077	0.8182	0.8791
False Alarm P_{fa}	0.1923	0.1818	0.1209

latter can only be used for classification.

In Figure 7, we further demonstrate the specific results of the test. Figure 7(a) - 7(d) are the test results of YOLO V3, and Figure 7(e) - 7(h) are the test results of YOLO V2. For some simple tasks, YOLO V3 and V2 are detected in addition to the boat, but the positioning of the V3 is more accurate, as shown in the Figure 7(a) and Figure 7(e). For the Figure 2(a) and Figure 2(b), YOLO V3 correctly detected the boat, but V2 did not.

IV. CONCLUSION

Through a large number of experiments, this paper proposes a sea surface small boat detection method based on YOLO V3 network for Radar data set. The result shows that this method can distinguish the spectrum between the small boat and sea clutter, and achieves high prediction accuracy.

REFERENCES

[1] P. Shui , D. Li and S. Xu, "Tri-feature-based detection of floating small targets in sea clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 2, pp. 1416-1430, 2014.

[2] N. Su , X. Chen , J. Guan and Y. Li, "Deep CNN-based Radar Detection for Real Maritime Target under Different Sea States and Polarizations," in *International Conference on Cognitive Systems and Information Processing*. IEEE, 2018.

[3] Y. Lecun , L. Bottou , Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[4] J. Redmon and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection". *Available online at http://arxiv.org/abs/1506.02640*, 2015.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger". *Available online at http://arxiv.org/abs/1612.08242*, 2016.

[6] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement". *Available online at http://arxiv.org/abs/1804.02767*, 2018.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on October 11,2020 at 03:48:12 UTC from IEEE Xplore. Restrictions apply.