A real-time detection USV algorithm based on bounding box regression

Yihong Zhang, Sen Wu *, ZiHao Liu, Yijin Yang, Di Zhu and Qian Chen

College of Information Science and Technology, Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, DongHua University Shanghai, China
2181328@mail.dhu.edu.cn

Abstract. With the rapid development of unmanned surface vessels (USV) applications, USV are widely used in military intelligence collection, target monitoring, and shipping services. However, this also brings security issues to shipping and the country. For more efficient detection of USV, in this paper, an enclosing center distance Intersection over Union (E-ClIoU) algorithm is proposed for real-time USV detection, where the normalized distance between center points of the smallest enclosing box and target box is combined with bounding box regression. Meanwhile, we designed a new USV data set. The extensive experiments on USV data set have demonstrated that the proposed approach has achieved better convergence speed during training and a significant accuracy prediction.

1. Introduction
As a new type of maritime ship, USV have been widely used in information collection, marine archeology, meteorological survey, resource exploration, maritime search and rescue, ship escort and anti-submarine work, which save a large amount of human resources and ensure personal safety. However, as an emerging field, USV have risks that can’t be controlled with existing resources, such as the passage of unmanned ships in restricted sea areas, illegal mapping around islands and reefs, and threats to normal shipping safety. So USV should be carefully maneuvered and positioned to solve the existing problems. Scholars have spent a lot of effort and time in evading and identifying USV, and the technology of detecting ships based on radar wave technology has been applied [1]. However, due to the complicated navigation environment of USV and the influence of buoys, this method is not universal. With the widespread popularity of airborne cameras and the huge increase in pixels, it is possible to identify USV through pictures using computer vision technology [2].

USV have the characteristics of high speed, small size and strong maneuverability. At the same time, because the imaging effect of the sea surface target is easily affected by the complex marine environment, such as weather, background clutter and other factors. these have proposed more detection capabilities of the algorithm High requirements.

Traditional manual feature-based detection algorithms not only require a large amount of training data, but the generalization ability of the algorithm is also affected by the parameters that need the technical person to adjust, which requires certain prior experience. Typical algorithms have HOG, Fast and DPM [3-5]. The emergence of deep convolutional neural networks, especially the two-stage algorithm Faster R-CNN [6] based on region proposals and the one-stage algorithm YOLO [7] based on regression are now widely used in pedestrian and vehicle detection. In recent years, scholars have made some improvements on the anchor box in order to frame the object position more accurately and
improve the detection accuracy of the algorithm. This provides more technical support based on real-
time speed to improve detection accuracy.

In this paper, we re-discuss the anchor box that is sampled uniformly over the spatial domain with a
predefined set of scales and design a new algorithm to regress the target area. It has been proved by
experiments that the improved box regression strategy has a greatly improved on the detection
accuracy of USV.

In Section 2, we introduce the popular methods of USV detection and the main methods of
bounding box regression. In Section 3, The proposed approach is explained. The experimental results
are shown in Section 4 based on the proposed algorithm. In Section 5, we summarize our approach
and propose directions for future work.

2. Related Work

2.1. USV Detection
The ocean contains abundant resources. In recent years, more and more USV operating equipment has
been applied to various tasks. Researchers have attempted used different methods to effectively
identify and accurately locate these USV, mainly the identification methods are radar detection and
radio detection. Liu et al. [8] proposed the USV-based laser fluorosensor system to detect and control
USV, where mainly used wireless communication technology. Ji et al. [9] proposed the radar image
target detection algorithms for embedded marine radar target detection system, where suitable for
objects recognition of surface of sea.

2.2. Bounding Box Regression
Detection and recognition are the basic problems in image field. Target detection not only implements
the classification of targets, but also solves the problem of target positioning, that is, obtaining the
position information of targets in the original image. Bounding box regression is the most basic
module in visual tasks. In order to obtain accurate positioning results, bounding box regression is all
applied regardless of target detection, target tracking or instance segmentation.

Most of the bounding box regression loss adopts regression loss. For example, the typical one-stage
detection algorithm YOLO uses mean square error (MSE) loss to regress the coordinates of the
bounding box, where the scale of predicted bounding box is adjusted by regressing parameters which
is square roots for \( w \) and \( h \). In two-stage detection algorithm R-CNN [10], one selective search
algorithm is introduced to extract prior bounding box from the input image that is extracted image
features by backbone network. Then, the size offset can be achieved by comparing the predicted
bounding box and the prior bounding box. Finally, a L2 loss is used to optimize the objective. Not
long after, a smooth-L1 loss is proposed by Fast R-CNN [11] to prevent the explosion phenomenon of
gradient. Focal loss [12] is modified based on the cross-entropy loss function to solves the problem of
serious imbalance between positive and negative samples in object detection, which is mainly used for
multi-class regression problems. In the one-stage target detection algorithm, training will be
dominated by a large number of negative samples. This paper only identifies the two objects category
problem, so we don’t consider the dense prior bounding box to regress the bounding box locations and
sizes.

IoU loss [13] is first proposed for face detection. Compared with the traditional softmax function, it
can be seen that directly optimizing IoU can lead to better performance, which provides a new
technical route for multi-class detection problems. GIoU loss [14] is proposed to solve the situation
where the IoU loss gradient disappears because the prediction bounding box and the target bounding
box do not overlap, but there are still problems of slow convergence and inaccurate regression. For
merging the normalized distance and IoU, Zheng et al. [15] propose DIoU loss, which can quickly
lock the best anchor box, but there are redundant calculations when the most suitable box is selected to
regress the target box. In this paper, we proposed a reasonable improvement to address the weakness
of DIoU and prove the advisability of our algorithm by calculating the USV accuracy of detection on specific epoch. In the next part, we will also exhibit explanation from the principle of algorithm.

3. Our Approach
This paper proposes an accurate positioning detection USV algorithm E-CIoU. By modeling the most representative one-stage detection YOLOv3 algorithm bounding box and redefining the loss function to assist real-time operations, the accuracy of detection results is improved. Multi-level fusion prediction using a hierarchical structure similar to the feature pyramid increases the possibility of detecting small targets as much as possible. Taking the bounding box into account of the regression loss, a loss function similar to the mean square error is used to predict the score for each target box, and the position accuracy is improved while maintaining the accuracy of the detection category. In addition, although only the binary classification problem is considered in this article, in terms of the versatility of the model, little change is implemented when detecting multiple classification problems.

3.1. A Network Architecture
The backbone part of the network uses darknet-53 structure. Compared with other backbone networks, darknet-53 has a greater advantage in classification accuracy. However, this article focuses on location accuracy, and continues to use logistic regression to predict the score of each category on the classification problem. The input image is scaled to fix the image size to a 608 * 608 tensor, and then the feature extraction operation is completed through 52 convolutional layers. The three-level prediction structure of the backbone network divides the image into 76 * 76, 38 * 38, and 19 * 19 tensors for prediction. The low-level convolutional receptive field is small, so which is responsible for detecting small targets. This level will dominate when the proportion of USV in the image is relatively small. The receptive field of deep convolution is large which lead to contain more high-level semantic information, the task of detecting large targets will take place here. When the proportion of USV in the image is relatively large, this level will dominate.

3.2. Loss Function
Intersection over Union (IoU) is the most commonly used indicator in detection tasks, where the prediction accuracy is explained by the ratio of the intersection and union of the prediction box and the ground truth box. Because IoU is a concept of ratio, the bigger ratio the more accuracy, it is insensitive to the scale of the target object. In the YOLO target detection algorithm, the accuracy of the detection is determined by setting the IoU threshold. The obvious insufficiency is that it can’t accurately locate the object. Repeated judgement for this process consumes a lot of computing resources and detection time. Generalized Intersection over Union (GIoU) proposed a method that IoU is designed as a distance measure, which is similar as L2 norm can be used to calculate the regression loss. Suppose we now have the predicted bounding box \( B^p \) and ground truth bounding box \( B^gt \) coordinates, respectively \( B^p = (x_1, y_1, x_2, y_2) \), \( B^gt = (x_1^gt, y_1^gt, x_2^gt, y_2^gt) \), where \( x_1 < x_2 \), \( y_1 < y_2 \), we can calculate area of \( B^p \) and \( B^gt \) by the following formula:

\[
\mathcal{A}^p = (x_2 - x_1) \times (y_2 - y_1)
\]

Where \( \mathcal{A}^p \) is the area of the predicted bounding box, it represents size of the target box which is given by the network.

\[
\mathcal{A}^gt = (x_2^gt - x_1^gt) \times (y_2^gt - y_1^gt)
\]

Where \( \mathcal{A}^gt \) is the area of the ground truth bounding box, it represents size of the target box which is actually exist by labeling in the picture. \( \mathcal{I} \) is the intersection area of \( B^p \) and \( B^gt \), it is shown in equation 3.

\[
\mathcal{I} = \left( \min\left( x_2, x_2^gt \right) - \max\left( x_1, x_1^gt \right) \right) \times \left( \min\left( y_2, y_2^gt \right) - \max\left( y_1, y_1^gt \right) \right)
\]
If \( \min\left(x_1, x_1^{\prime}\right) < \max\left(x_1, x_1^{\prime}\right) \) or \( \min\left(y_1, y_1^{\prime}\right) < \max\left(y_1, y_1^{\prime}\right) \), \( \mathcal{I} \) should be set as 0. It means that \( B^p \) and \( B^{\prime} \) haven’t the common part. \( \mathcal{U} \) is the union area of \( B^p \) and \( B^{\prime} \).

\[
\mathcal{U} = \mathcal{A}^p + \mathcal{A}^{\prime} - \mathcal{I}
\]

So, we get IoU loss as

\[
\mathcal{L}_{\text{IoU}} = 1 - \text{IoU} = 1 - \frac{\mathcal{I}}{\mathcal{U}}
\]

It is easy to find \( \mathcal{L}_{\text{IoU}} \) is faulty that the gradient is not falling or falling into a local optimum when \( \mathcal{I} \) equals zero. The most direct consequence is not getting the desired training model.

Zheng et al. [15] proposed CIoU loss to accelerate calculation and solved the problem of gradient vanishing. CIoU loss can be expressed by the following equation.

\[
\mathcal{L}_{\text{CIoU}} = 1 - \text{IoU} + \frac{d^2}{c^2} + \alpha v
\]

Where \( d \) is the distance of center points of \( B^p \) and \( B^{\prime} \), \( c \) is the diagonal length of the smallest enclosing box \( B^p \) which covering \( B \) and \( B^{\prime} \) box. \( \alpha \) is a trade-off parameter which always is positive. \( v \) measures the consistency of aspect ratio,

\[
v = \frac{4}{\pi^2} \left( \arctan \frac{w^{\prime}}{h^{\prime}} - \arctan \frac{w}{h} \right)^2
\]

The trade-off parameter \( \alpha \) is defined as

\[
\alpha = \frac{v}{(1 - \text{IoU}) + v}
\]

We can find parameter \( \alpha \) is restricted by IoU that is an insensitive variable. So, in order to make the best of the impact of \( v \) on loss, we consider take some improvement according to the characteristic of spatial domain target and predicted bounding box. It can be shown that every component in the algorithm of table 1 has feasibility.

**Table 1. The core algorithm structure of bounding box regression.**

**Algorithm:** E-CIoU as bounding box losses

**Input:** Predicted \( B \) and ground truth \( B^{\prime} \) bounding box coordinates:

\[
B = (x_1, y_1, x_2, y_2), \quad B^{\prime} = (x_1^{\prime}, y_1^{\prime}, x_2^{\prime}, y_2^{\prime}), \quad \text{where} \quad x_1 < x_2, \quad y_1 < y_2.
\]

**Output:** \( \mathcal{L}_{E-CIoU} \).

1. Calculating center points of \( B^{\prime} \) : \( o_x^{\prime} = \frac{x_1^{\prime} + x_2^{\prime}}{2}, \quad o_y^{\prime} = \frac{y_1^{\prime} + y_2^{\prime}}{2} \).

2. Calculating center points of the smallest enclosing box \( B^p \) which covering \( B \) and \( B^{\prime} \) box:

\[
x_1^p = \min\left(x_1, x_2\right), \quad x_2^p = \max\left(x_1, x_2\right), \quad y_1^p = \min\left(y_1, y_2\right), \quad y_2^p = \max\left(y_1, y_2\right),
\]

\[
o_x^p = \frac{x_1^p + x_2^p}{2}, \quad o_y^p = \frac{y_1^p + y_2^p}{2}.
\]

3. Calculating the distance of center points of \( B^p \) and \( B^{\prime} \) : \( d_1 = \sqrt{\left(o_x^{\prime} - o_x^p\right)^2 + \left(o_y^{\prime} - o_y^p\right)^2} \).

4. Obtaining trade-off parameter \( \beta \) : \( \beta = \frac{d_1^2 + 10^{-16}}{d_2^2 + 10^{-16}} \).

5. E-CIoU = \( \text{IoU} - \frac{d^2}{c^2} + \beta v \)

6. \( \mathcal{L}_{E-CIoU} = 1 - \text{E-CIoU} \).
Value of IoU will be set as 0 when prediction box and target box don’t have intersection area, therefore, $\alpha$ will have a very high proportion as a trade-off parameter, which will confuse the direction of the gradient descent of loss. Because it is immanent shortage to select error anchor box for the regression of aspect ratio. As shown in figure 1, we utilize the ratio of square of center distance as a trade-off parameter for reducing the defect of inappropriate anchor box to regress aspect ratio. $d_1$ is always less than or equal to $d$. In the cases we listed, when the two boxes have intersection or no intersectional area, $\beta$ will be very small, which is situation we want. $\beta$ is set as 1 and $\beta_v$ returns to the maximum when the anchor box contains the target box, which means predicted box based on the regression of the specific anchor box. DIoU [15] will dominate when the target box contains all anchor boxes smaller than its size, but our approach can also work when $d$ is tiny. It is worth noting that the gradient explosion will occur when the denominator $d$ of $\beta$ is 0 to calculate the direction of gradient. So $10^{-16}$ as a canonical value is added to $\beta$, the influence of which for the loss can be neglected under the premise of ensuring that the gradient direction is correct.

Figure 1. E-CIoU loss for bounding box regression. With the number of training times increasing, the normalized distance between center points of prediction box and target box can be minimized directly. The black box represents anchor box, $d$ is the distance of center point of it and target box. $c$ is diagonal length of the smallest enclosing box surrounding the two boxes. $d_1$ is the distance of center points of enclosing box and target box, where $d$ and $d_1$ are fused as a parameter in our approach.

Figure 2. Simulation experiment. regression error sum curves of different loss functions at iteration. Compared with other loss functions, E-CIoU has more quickly convergence speed and smaller error.
For verifying the rationality of this algorithm, our approach is added to the simulation experiment proposed by Zheng et al. [15] Compared with other loss functions, in figure 2, the experimental results show that our approach has more superiority in the 1,715,000 regression cases.

3.3. Production of USV dataset
Data sets are the basis for testing algorithm performance in practical problems. Commonly used data sets in the field of computer vision include cifar-10, COCO, PASCAL VOC, etc. These data sets are used to test the algorithm's fitting ability and the comprehensive performance of the extracted information. In this paper, a data set is proposed to deal with practical problems. USV are a kind of water surface objects. Considering the unity of the background, we select pictures from multiple angles and add some experimental data pictures to ensure the generalization performance of the sample while diversity of the sample. At the same time, due to the inconsistency of the scale, it is difficult to distinguish ships such as yachts that have similar characteristics to USV. In the data set, we actively exclude the interference of such pictures. In multi-class detection tasks, this effect can be effectively distinguished by providing enough training samples. In this application, only two-category problems are considered and the data set is kept as simple as possible, which also relieves the excessively high hardware requirements during the training process. The dataset of this article is labeled with LabelImg software, and the data format is similar to PASCAL VOC.

4. Results
The methods proposed and compared in this paper are implemented on the GPU of a single GTX1660ti. The detection effect of USV are shown in figure 3. By comparing the four pictures shown, compared with the latest methods, we can ensure that the position accuracy of the calibration target position box is significantly improved on the premise of ensuring accurate classification, especially under a single USV. The detection result of the algorithm is given in Table 2. When the detector is set 0.5 for IoU threshold, AP50 is the Average Precision (AP) of the algorithm, which shows that E-CIoU and CIoU losses are better than the classical DIoU loss. The three-level fusion detection method of YOLO has superior in detecting small targets, the AP75 data shows that the E-CIoU loss regression method is about 7.3% promotion in the precision, which means more accurate than the CIoU. Even the threshold is set 0.9, the detection performance of E-CIoU still improves by about 2.7%.

Table 2. Quantitative comparison of E-CIoU with other multiple methods of IoU losses on the test set.

| Loss name | Image size | Boxes_AP50 | Boxes_AP75 | Boxes_AP90 |
|-----------|------------|------------|------------|------------|
| DIoU      | 608        | 80.06      | 64.37      | 2.69       |
| CIoU      | 608        | 87.31      | 68.14      | 7.27       |
| E-CIoU    | 608        | 87.85      | 75.71      | 9.96       |

In the comparison of the second and third pictures in figure 3, we notice that after the detection algorithm is adjusted in the regression pattern, except the object border is based on the accurate calibration of the actual experimental picture data, the left and right of side of the chord side of the USV borders are also adjusted. In addition to the experimental data, the test set of this article also added some commercial USV test pictures. In the comparison of the fourth picture, for single USV detection, this paper can reduce the size of the detection border to the extreme. Under multiple backgrounds, such as the interference of a manned yacht, the E-CIoU detection method will also show a certain detection deviation under the premise of ensuring the detection of the picture, as shown in the first comparison picture. Based on this kind of problem, it will be left for future work. For the objectivity and fairness of the results, all the above comparison results are implemented in the same hardware equipment and the same training parameters.
Figure 3. These pictures are used to demonstrate the superiority of the proposed method that by comparing the results of using CIoU loss function.

The main contribution of this paper is to make a more scientific adjustment of the regression method of the bounding box, and combined with YOLO's multi-scale feature extraction backbone network for USV detection. With the development of video technology and the continuous expansion of data set, the recognition accuracy and generalization ability of data will continue to improve. In recent years, the improvement of communication speed, especially the application of 5G technology, the real-time detection performance of the algorithm can be used to identify ships at sea or offshore, which can greatly reduce the dependence of personnel. The airborne camera can effectively reduce demand in the facility configuration and just need some personnel costs. Compared with all current biological surveillance technologies, the algorithm is continuously monitored through computer technology. People only need to perform secondary recognition and detection results, which can effectively reduce missed inspections caused by human negligence. The residual network structure of darknet can effectively extract image features. In view of the regression method has been improved in this paper, only some network configurations need to be adjusted, the algorithm can more accurately identify more objects.

5. Conclusion/Future Work
In this paper, we propose the E-CIoU loss for bounding box regression. By directly minimizing the center point of the prediction box and the target box, the standardized distance between the smallest enclosing box and the target box is incorporated in bounding box regression to make the speed and performance of convergence of network more superior. The purpose of the USV data set is to complete the design of the YOLO based USV detection algorithm. Compared with the latest DIoU and CIoU losses, the E-CIoU loss has a significant improvement in detection accuracy. The proposed loss function can be easily integrated into other object detection frameworks. In future work, we will conduct comparative tests in other detection frameworks to find better results. Due to the limited size of the data set, we will continue to expand the number of data sets and more complex scenes for widespread use.

Acknowledgments
This research was funded by [the Fundamental Research Funding for the Central Universities of Ministry of Education of China] grant number [2232018A3-03] and [the Special Project Funding for the Shanghai Municipal Commission of Economy and Information Civil-Military Inosculcation Project]
“Big Data Management System of UAVs” grant number [JMRH-2018-1042]. The author especially thanks the participants of this project for their pertinent advice and help.

References
[1] Zhuang, J. Y., Zhang, L., Zhao, S. Q., Cao, J., Wang, B., and Sun, H. B. (2016) Radar-based collision avoidance for unmanned surface vehicles. China Ocean Engineering, 30(6), 867-883.
[2] Luo, S., Singh, Y., Yang, H., Bae, J. H., Dietz, J. E., Diao, X., and Min, B. C. (2019) Image Processing and Model-Based Spill Coverage Path Planning for Unmanned Surface Vehicles. In: OCEANS 2019 MTS/IEEE SEATTLE. Seattle. pp. 27-31.
[3] Dalal, N., and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego. pp. 886-893.
[4] Dollar, P., Appel, R., Belongie, S., and Perona, P. (2014) Fast Feature Pyramids for Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(8), 1532–1545.
[5] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010) Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1627–1645.
[6] Ren, S., He, K., Girshick, R., and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149.
[7] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR. Seattle. pp. 779-788.
[8] Liu, D., Luan, X., Zhang, F., Jin, J., Guo, J., and Zheng, R. (2016) An USV-based laser fluorosensor for oil spill detection. In: 10th International Conference on Sensing Technology, ICST. Nanjing. pp. 1-4.
[9] Ji, X., Zhuang, J. Y., and Su, Y. M. (2014) Marine Radar Target Detection for USV. Advanced Materials Research, 1006–1007, 863–869.
[10] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus. pp. 580-587.
[11] Girshick, R. (2015) Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision. Santiago. pp. 1440-1448.
[12] Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, Piotr. (2020) Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 318-327.
[13] Yu, J., Jiang, Y., Wang, Z., Cao, Z., and Huang, T. (2016) Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia. New York. pp. 516-520.
[14] Rezatofighi, H., Tsoi, N., Gwak, J. Y., Sadeghian, A., Reid, I., and Savarese, S. (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 658-666.
[15] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2019) Distance-iou loss: faster and better learning for bounding box regression. arXiv preprint arXiv:1911.08287.