International Journal of Intelligent Robotics and Applications
https://doi.org/10.1007/s41315-022-00238-y

REGULAR PAPER

LEDet: localization estimation detector with data augmentation for ship detection based on unmanned surface vehicle

Yang Zhou[1,2] · Jingling Lv[1,2] · Yueying Wang[1,2] · Chang Liu[1,2] · Songyi Zhong[1,2] · Guozhu Tan[1,2] · Jiacheng Sun[1,2]

Received: 5 November 2021 / Accepted: 28 April 2022

Abstract
Ship detection is significant for monitoring ports, especially contributing to the safe driving of Unmanned Surface Vehicle (USV). However, recent ship detection based on deep learning lacks complete ship datasets and uses the classification score as the ranking basis, which harms their performance. To address the problems, we present a one-stage localization estimation detector (LEDet) with ship-customized data augmentation. Specifically, we integrate the localization quality estimation into the classification branch as a soft label localization score. We further apply ship-customized data augmentation named "cutting-transform-paste" to expand ship datasets without manual annotation. Hence, a large number of diverse ship datasets can be created. Extensive experiments show that our LEDet consistently exceeds the strong baseline by 8.0% COCO-style Average Precision (AP) with ResNet-50. It significantly improves the ship detection performance.

Keywords Deep learning · Object detection · Ship detection · USV · Autonomous systems

1 Introduction

Unmanned Surface Vehicle (USV) is widely used in military and civil ship detection. USV is a kind of surface robot without manual operation. Hence, the perception technology is significant for the safe driving of USV. As the core of the unmanned system, USV perception can be applied in the port terminal to automatically detect ships and maintain marine safety. In recent years, USV is used to detect illegal ships such as containerships, aircraft carriers, and so on with the infrared camera. Therefore, the desired detector must consider how to improve the performance of the USV perception. However, two problems existed in the optical ship detection methods lack high-quality datasets containing multiple types of ships, and inaccurate location estimates.

Massive datasets are crucial to the success of any machine learning task, and the task of ship detection is no exception. Unfortunately, there are no complete datasets for ship detection, and creating one is time-consuming. For example, the popular large datasets such as MS COCO (Lin et al. 2014) and PASCAL VOC classify ships as only one kind and contain only 4814 and 300 unclassified ship images, respectively. In a specific application, classifying ships into multiple categories is necessary. Consequently, we collect the dataset of ship images and annotate 6 kinds of ship categories in this research. But labeling is a tough job. For example, COCO required 2000 working hours to create, and 22 workers' hours were spent per 1000 instance masks. This labor-intensive process means it is imperative to develop more efficient methods of creating datasets.

A simple method for generating ship image datasets proceeds as follows: annotate a portion of single datasets, then paste the labeled ships into the new sea pictures. With the random reorganization of image data, many ship images are generated. To some extent, this method can also create much more ships.

Moving from the problem of datasets to accurate detection, we must consider the cause of unreliable ship detection: the classification scores are used as the metric of Non-Maximum Suppression (NMS) (Neubeck and Gool 2006). However, the classification scores cannot fully represent the ranking metric. Consequently, the algorithm will select unreliable detection results, as shown in Fig. 1.

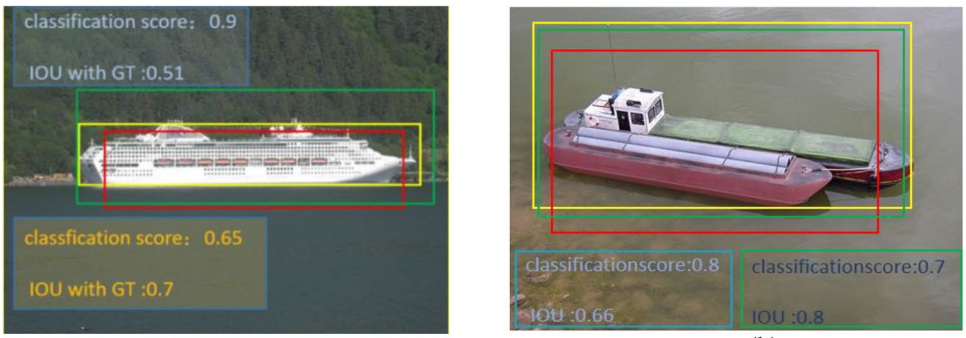
To estimate localization quality for selecting reliable detection results, we consider integrating the ship location

✉ Songyi Zhong
zhongsongyi@shu.edu.cn

[1] Shanghai University, Shanghai, China

[2] Engineering Research Center of Unmanned Intelligence Marine Equipment, Ministry of Education, Shanghai, China

Published online: 26 May 2022



(a) (b)

Fig. 1 The misalignment between ship classification confidence and ship localization accuracy is illustrated in the cases. The yellow bounding boxes denote the ground truth (GT), while the red and green bounding boxes are both detection results yielded by Feature Pyramid Networks (FPN) (Lin et al. 2017). As shown in picture (**a**), the classification score of the red bounding box is 0.65 and the green bounding box is equal to 0.9. IoU between the red bounding box and GT is 0.7, and IoU between the green bounding box and GT is 0.51. Using classification confidence as the ranking metric will cause accurately bounding boxes (in red) incorrectly eliminated in the traditional NMS procedure. Similarly, the green bounding box with a lower classification score but higher IoU with GT will be removed in picture (**b**)

information into the classification task to predict the class-localization score. Finally, we can exploit the prediction score called localization quality score as localization accuracy to select detections in NMS.

Based on these ideas, we design a novel ship detection method named LEDet with data augmentation. There are two parts of the detector. One is a simple, yet effective ship-customized data augmentation method named "cutting-transform-paste" for expanding and strengthening ship datasets. The other part contains a localization estimation network to predict the class-localization score. By combining the two ideas, LEDet improves ship detection quality.

Contributions: to summarize, our contributions are as follows:

1. We put forward a novel ship detection method (LEDet with data augmentation) for USV that accurately detects ship targets.
2. We collect a large amount of multi-category ships and annotate each ship for constructing an initial ship dataset. To rapidly augment datasets, we propose a simple method "cutting-transform-paste" to generate complex and diverse ship datasets.
3. We propose a novel localization estimation detector (LEDet) that addresses the estimation of localization accuracy. It is applied in the NMS procedure to accurately rank a large number of candidate bounding boxes.
4. We achieve 62.7%AP on our ship datasets. Our detector consistently exceeds the strong baseline RetinaNet (54.7% AP) by 8.0%AP with the ResNet-50 (He et al. 2016) backbone from extensive experiments.

The remainder of this paper is organized as follows. Section 2 reviews the related work. In Sect. 3, we propose our novel data augmentation method and LEDet. All experiments are given in Sect. 4. In the last section, we conclude this paper.

2 Related work

At present, traditional ship detectors use two methods to identify ships: ship structure and the characteristics of ships, or threshold-based edge detection respectively. The former ship structure and shape are used for manual feature design, while the latter directly utilizes the threshold to detect ship edge features. In 2012, Fefilatyev et al. presented a novel algorithm for the open sea. The large datasets collected from a prototype system (Fefilatyev et al. 2012) achieve the ship detection precision of 88%. In 2017, Zhang et al. proposed a new ship target detection algorithm for visual maritime surveillance. The three main steps: horizon detection, backbone modeling, and backbone subtraction, are all based on the discrete cosine transform (Zhang et al. 2017).

Although traditional methods have achieved good results, there are some challenges. For example, traditional methods (Chen et al. 2017; Zhi et al. 2014; Fingas and Brown 2014) detect after sea-land segmentation and utilize the hand-crafted features for discrimination. But they have poor performance in near-shore areas and have difficulty ruling out false alarms. Additionally, these methods suffer under real-world conditions like noise, complex backgrounds, and minute hull differences.

With the rapid development of the deep convolutional neural network (CNN) in recent years, CNN has made great progress in the application of human faces, pedestrians, and other scenes. Besides, CNN is far superior to traditional methods in accuracy and speed. At the same time, there are

some ship detection algorithms such as the YOLO series (Redmon and Farhadi 2017, 2018; Bochkovskiy et al. 2020). Nowadays, we can simply classify the objection detection methods into anchor-based and anchor-free detectors by whether to use anchors. The anchor-based detectors contain two-stage and one-stage methods. Moreover, the anchor-free methods divide into keypoint-based and center-based methods. (1) Anchor-based detectors: an anchor-based approach starts to get regional proposals with different but fixed scales and shapes by placing a large number of anchors. These anchors are then considered as object proposals and an individual classifier is trained to determine the objectness as well as the class of each proposal such as the two-stage detector Faster RCNN (Ren et al. 2017). (2) Anchor-free detectors: an anchor-free approach does not assume the objects to come from uniformly distributed anchors. Recently, anchor-free approaches have been greatly promoted. Keypoint-based detectors such as CornerNet (Law and Cornernet 2018), CenterNet (Duan et al. 2019), ExtremeNet (Zhou et al. 2019), etc., group more than one keypoints into an object, while the Center-based detectors such as FCOS (Tian et al. 2019), FoveaBox (Kong et al. 2020), SAPD (Zhu et al. 2020), etc.., however, these algorithms based on CNN require larger datasets, which indeed emphasize the significance to build an adequate ship dataset for our specific detection task.

2.1 Boat datasets and data augmentations

Current datasets contain a few ship images and a limited range of image types. There are currently only six well-annotated ship datasets, as shown in Fig. 2. Among these datasets, HRSC2016 is annotated by Kaggle. NWPU VHR-10 annotated by Northwestern Polytechnic University has only 800 images in total, including 650 targets of ships a pitifully small number for deep learning. Among currently existing ship datasets, they are almost made up of exclusively remote sensing images and few ship datasets based on RGB images. A complete multi-class ship dataset based on RGB images has more refined features than remote sensing images and contains higher resolution. Hence, based on RGB ship datasets are beneficial for small target detection and localization quality.

For data augmentations, there are some works for detection, including general object detection (Zoph et al. 2020; Chen et al. 2021a) and specific object detection like pedestrian detection (Tang et al. 2021; Chen et al. 2021b). Unfortunately, data augmentation receives little attention in the computer vision community. There is a much greater volume of work on backbone architectures (He et al. 2016; Krizhevsky et al. 2012) and detection/segmentation frameworks (Girshick 2015; Girshick et al. 2013; He et al. 2017). Despite this lack of attention, data augmentations such as random crop (Krizhevsky et al. 2012; Simonyan and Zisserman 2014; Lecun and Bottou 1998; Szegedy et al. 2015), color jittering (Szegedy et al. 2015), AutoRandAugment (Cubuk et al. 2019, 2020) have played a crucial role in achieving state-of-the-art results on image classification (He et al. 2016) and self-supervised learning (He et al. 2020).

These methods are universally and mainly used for encoding invariances to data transformations. Data augmentation generates more diverse data so that the network can learn more general features and improve its generalization performance.

2.2 Estimation of localization quality

Accurate estimation of localization quality is a challenging and important topic. Many methods have been proposed. In traditional methods, duplication detection is suppressed while ignoring localization accuracy and the classification scores are typically used as the metric for ranking the

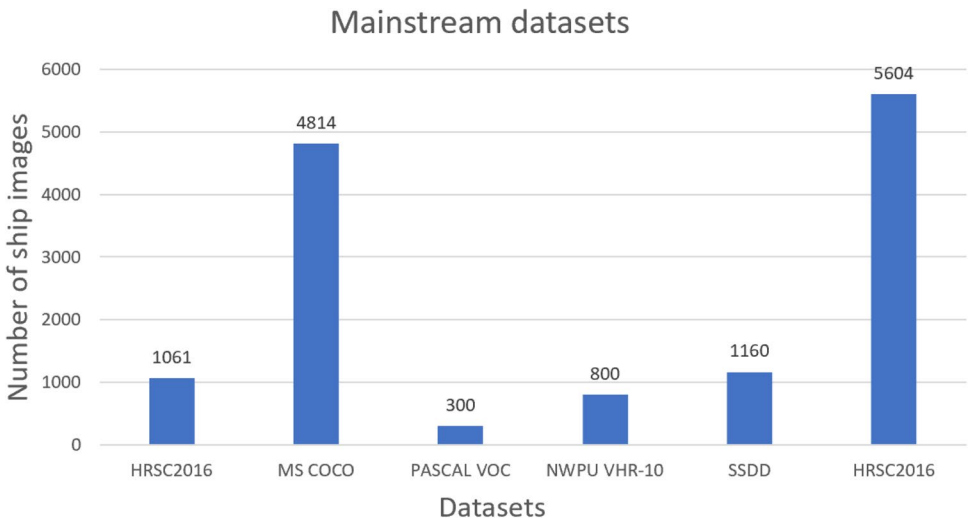


Fig. 2 Illustration of current mainstream datasets and the number of ship images in datasets

proposals. To estimate localization accuracy, IoU-Net (Jiang et al. 2018) and IoU-aware (Wu et al. 2020) use an additional network to predict Intersection over Union (IoU) as the representation of localization accuracy and exploit it to rank bounding boxes in NMS. FCOS ranks detections with predicted centerness scores and suppresses the low-quality detections. They all utilize an additional branch to perform localization quality estimation in a form of IoU or centerness scores. But this separate formulation causes two problems. First, separately formulating can cause inconsistency between training and testing as well as unreliable quality predictions mentioned in Generalized Focal loss (Li et al. 2020). The other is the inaccurate localization quality which is estimated by adding an additional branch. Because the negative samples do not participate in the training, it is inaccurate to use the score of extra branch prediction to estimate the localization quality.

To improve the inaccurate localization quality, IoU-balance (Wu et al. 2022) is proposed to assign different weights in the classification loss. They somewhat solve the inconsistency but do not solve it completely. Therefore, we propose a localization estimation detector (LEDet) to address the problem. In the classification task, we integrate ship localization information predicting the ship localization quality score to estimate localization accuracy. Finally, based on the ranking of localization quality score, we solve the problem of inconsistency between training and testing.

3 Method

3.1 Datasets generation

In this section, we mainly introduce how to construct the ship datasets. To build the ship detection datasets with precise annotation, we collect six categories of ships to construct initial ship datasets. The ship images were taken by both ordinary cameras and USV cameras to capture ship images and videos from multiple angles and different zoom ratios in the wharf and offshore waters. Specifically, in the process of taking, we use a variety of perspectives to take pictures of ships with different postures such as the bow, stern, and side of the ship. At the same time, we set different zoom ratios of the camera to take pictures with different resolutions. In practice, we set parameters such as 0.5, 1, 2, 5, and 10 times to obtain ship targets at different scales. In the end, we eliminate blurred pictures and keep the picture clearly with human subjective feelings. In addition, the images captured by USV do have some specific characteristics compared to usual images captured by RGB cameras on the shore or at close range, including small objects, different views of ships, unstable image quality influenced by the unique marine environment, and special view (horizontal to the sea level), etc. For some rare military ships such as aircraft carriers and submarines, we captured from public datasets such as ImageNet, PASCAL VOC, MS COCO and used LABELIMG software to annotate each picture. Next, we introduced the content details of the ship datasets. Ships are categorized as 'containership', 'sailing boat', 'aircraft carrier', 'speed boat', 'gondola', and 'submarine', as shown in Fig. 3. In the process of image feature extraction, some factors such as visibility scale, visual angle, illumination, background, and occlusion are fully considered. In the end, our datasets contain 17,932 ship images. Figure 4 and Table 1 show the detailed ship categories and ratios of different ship categories.

3.2 Data augmentation

To rapidly augment the number of images in the ship dataset, we present a new method named "cutting-transform-paste" (CTP).

The CTP method consists of three steps. In the cutting step, we cut and copy the ground-truth boxes of the ships with their category and location from each picture. In the transform step, many image transformation operations which contain flip, rotate, scaling, and so on are performed on the copied ship images. For each image transformation parameter, the guidelines we follow in our experiments are to facilitate image feature extraction. All parameters are the super parameters set by the author. To obtain ships of different sizes, we set a random scaling factor. If the size of the ship targets in the picture is less than 100×100, we enlarge the size of the ship by setting a $2\times$ scaling factor. Otherwise, we scale the size of a ship by $0.6\times$. The main purpose is to extract the different sizes of ships as long as they are clear. In addition, we found that the scaling factor cannot be set too large or too small and it will cause the ship to be unclear and affect the detection accuracy. The rotation angle is usually set to 45 degrees and 90 degrees. After these transformations, we paste these instances on the new sea images anywhere. All in all, the purpose of flip and rotation operations is to augment the perspectives of the ship. Other transformations such as block clipping and Gaussian noise are set to simulate ship shelter and the complexity of the sea background. Consequently, these parameters are determined according to the experimental results. Next, the transformed images are randomly pasted into new sea surface images that came from our team collected in advance, while generating new pictures. Consequently, the total number of ship datasets has been doubled. The general framework of the CTP structure is shown in Fig. 5.

By reorganizing the position, gesture, scale, and other relationships between the ship images and the sea scene, the diversity and complexity of the data are greatly expanded. 33,691 ship images could be obtained from our original

Fig. 3 Examples of constructed initial ship datasets and their categories. **a** containership **b** sailing boat **c** aircraft carrier **d** speedboat **e** gondola **f** submarine

17,932. Using this method makes it simple and effective to create new and challenging training pictures, which can improve detection performance.

When conducting the image transformation operations, there will be an infinite number of images generated theoretically. However, our experiments show that doubling the ship images is the best choice to get the best detection accuracy. Once we paste more than two times, the accuracy will decrease. Table 2 demonstrates this phenomenon.

3.3 Localization estimation network

To solve the inaccurate estimation of localization accuracy in ship object detection, we propose a localization

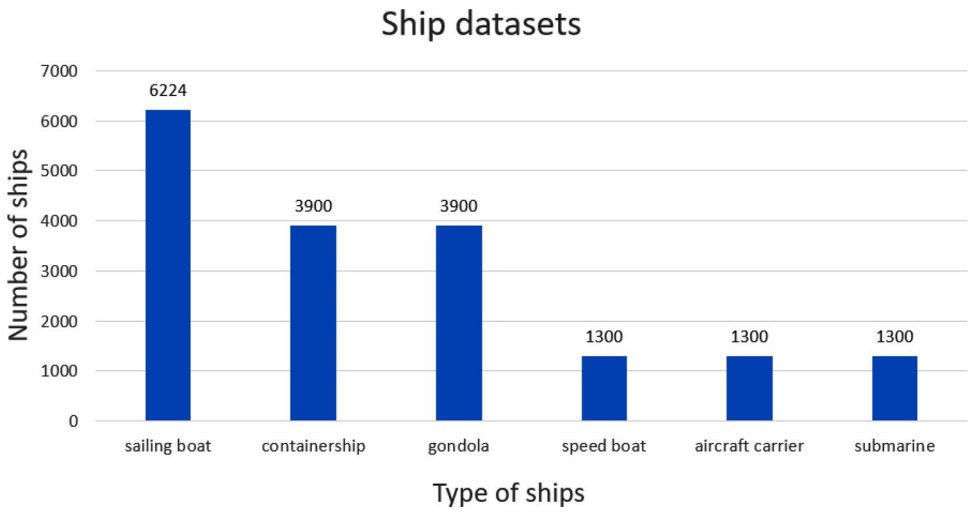


Fig. 4 Histogram statistics of the ship categories

Table 1 The detailed information of initial ship datasets

Category	Image numbers	Ratio
Sailing boat	6224	0.347:1
Containership	3900	0.218:1
Gondola	3900	0.218:1
Speed boat	1300	0.073:1
Aircraft carrier	1300	0.073:1
Submarine	1300	0.073:1

estimation detection network. In the localization estimation task, we smooth the original category label 0.1 and utilize the IoU between the predicted bounding box and GT as the localization label, as shown in Fig. 6. At the same time, the soft label localization score simultaneously represents the localization accuracy of bounding boxes. Therefore, we exploit the soft label localization score to address the representation of localization accuracy in one task.

Based on the motivation of the RetinaNet detector, we design a novel one-stage detector called LEDet. Instead of learning to predict the class label for a bounding box, we merge the localization accuracy score into the classification branch called localization quality score which can represent the certain class label and localization accuracy for a bounding box.

Figure 7 illustrates the network architecture of LEDet and the overall network architecture is the same as the RetinaNet.

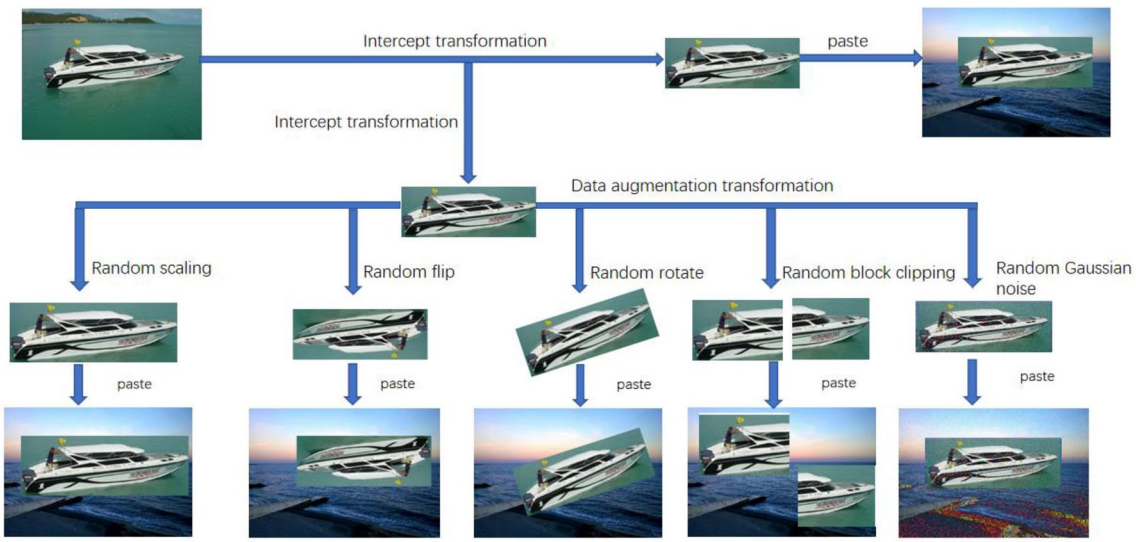


Fig. 5 Illustration of the proposed CTP structure. Firstly, we exploit intercept transformation to obtain the ship target image. Secondly, we employ the kinds of data augmentation transformations such as random scaling, random flip, random rotation, random block clipping, and random Gaussian noise to obtain challenge ship images. In the end, we paste the ship images into the new sea picture

Table 2 Compare the influence of doubling the ship images and tripling the ship images on detection accuracy

Method	AP	AP_{50}	AP_{75}
Doubling the ship images	62.7	84.6	68.6
Tripling the ship images	**58.9**	**81.3**	**64.0**

Bold values indicate our experimental conjecture

The whole network structure consists of three parts: backbone, neck, and detection head. The difference between the RetinaNet and the LEDet is the detector's head structure. The LEDet consists of two subnetworks. The bounding box regression subnet performs bounding box regression, while the localization score estimation subnet predicts the localization quality score. In the inference stage, the network outputs localization scores for six categories of ships. Hence, we exploit the highest score as its localization score and its certain ship label.

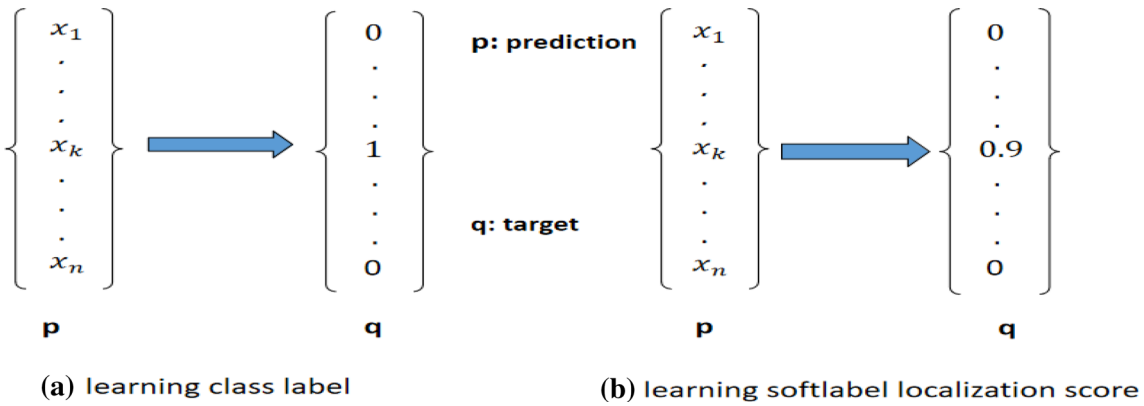


Fig. 6 An illustration of the localization estimation branch. Instead of learning to predict the class label for bounding box (**a**), we learn the soft label localization score (**b**)

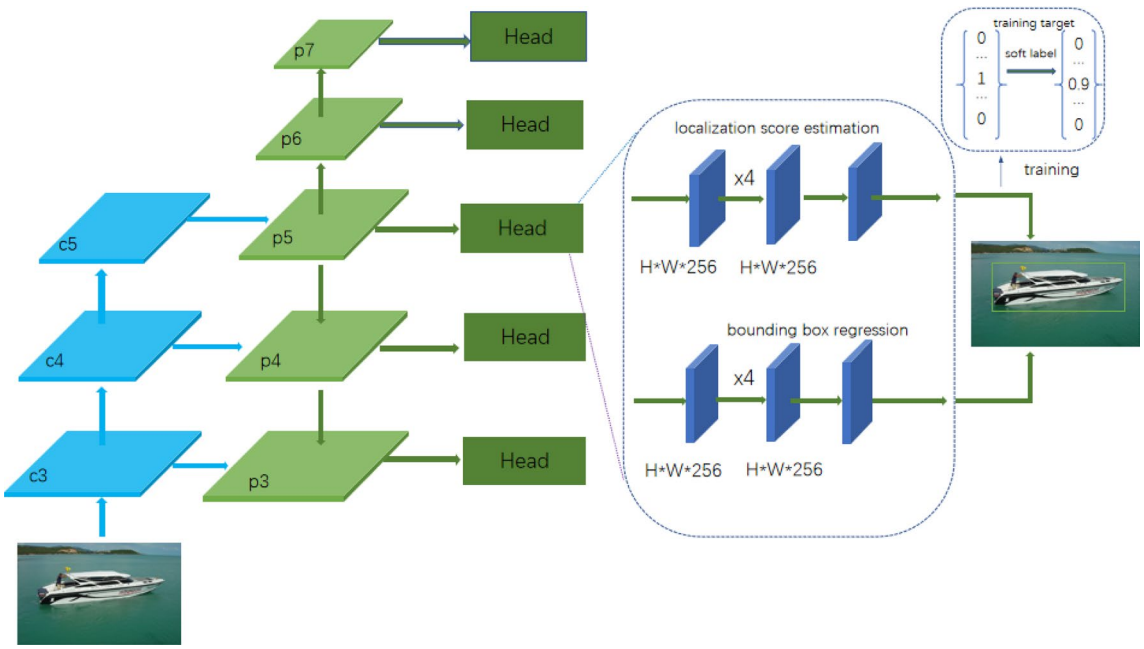


Fig. 7 The network architecture of our LEDet. The LEDet is built on the FPN(P3-P7). Its head consists of two subnetworks. One is for regressing the bounding box, and the other is for predicting the soft label localization quality score. H*W denotes the size of the feature map

3.4 Loss function and inference

The training of our LEDet is supervised by the soft loss.

$$\mathrm{Soft}(p,q)=\begin{cases}-q(q\log(p)+(1-q)\log(1-\mathrm{p})) & q>0\\ -\alpha p^{\gamma}\log(1-p) & q=0\end{cases} \quad (1)$$

where p is the predicted localization quality score, and q is the target IoU score. For a positive training example, q is set as the IoU between the generated boxes and ground truth (gt_IoU) bounding boxes. For the negative training example, the training target q for all classes is 0.

Our loss is designed for regressing the continuous localization score. Unlike focal loss which treats positives and negatives equally, our loss treats them asymmetrically. As Eq. (1) shows, we weight the positive example with the training target q. For negative examples, we reduce the loss contribution by set a factor of p^{γ}. But we do not down-weight positive examples in the same way. The main reason is that the positive samples are far less than the negative samples. Therefore, our loss focuses on high-quality positives rather than low-quality examples. Moreover, we need to balance the loss between positive samples and negative samples by setting an adjustable factor α.

3.5 Total loss function

Our total loss function is defined as follows.

$$\mathrm{Loss}=\frac{1}{N_{pos}}\sum_{i}\mathrm{Soft}\left(p_i,q_i\right)+\frac{\lambda}{N_{pos}}\sum_{i}q_iL_{bbox}\left(\mathrm{bbox_i},\mathrm{bbox_i^*}\right) \quad (2)$$

where p_i and q_i denote the predicted and gt_IoU at the location i on each level feature map of FPN, respectively. L_{bbox} is the GIoU loss (Rezatofighi et al. 2019) and the representation of the initial and ground-truth bounding box is $\mathrm{bbox_i}$ and $\mathrm{bbox_i^*}$, respectively. We weight the L_{bbox} with the training target q_i, which is a value ∈ (0, 1] for positive examples and 0 otherwise. λ is the balance weight for L_{bbox}, we empirically set λ = 1.5 in this paper. N_{pos} means the number of positive examples and is used to normalize the total loss.

We forward an input image through the network and remove redundant detections by using the NMS.

4 Experiments

4.1 Experiment settings

We evaluate the LEDet based on our augmented ship datasets. Specifically, the datasets contain 33,691 images with a resolution of 535 × 300. Among them, 29,169 images are used for training, and the rest are used for testing. We adopt the standard COCO-style Average Precision (AP) as the evaluation metric.

4.2 Implementation and training detail

We implement LEDet with MMdetection (Chen et al. 2019). Unless specified, we adopt the default hyper-parameters used in MMdetection. We use the ImageNet (Jia et al. 2009) pre-trained ResNet-50 with a 5-level feature pyramid structure as the backbone. During training, the input images are resized to keep their shorter side being 800 and their long side below 1334. In the ablation study, the networks are trained using the stochastic gradient descent (SGD) algorithm for 90 K interactions (denoted as 1 × schedule) with 0.9 momentum, 0.0001 weight decay, and 8 batch size. The initial learning rate is set as 0.005 and decayed by 0.1 at iterations 60 K and 80 K.

4.3 Inference detail

During the inference phase, we resize the input image in the same way as the training phase and then forward it through the whole network to output the predicted bounding boxes with a predicted localization quality score. We first filter out those bounding boxes with a 0.05 threshold and select at most 1 k top-scoring detections per FPN level. Then, the selected detections from all levels are merged and redundant detections are removed by NMS with a threshold of 0.5 to yield the final results.

5 Ablation study

To better understand the impact of each module, we investigate the performance of each module.

5.1 Soft label

We investigate the effect of the soft label as the training target in the classification subnet and research how the localization quality score we predicted affects the performance

Table 3 Comparison of RetinaNet and our proposed LEDet (without data augmentation) method in our initial ship datasets

Method	AP	AP_{50}	AP_{75}
RetinaNet	54.7	76.6	60.9
LEDet	**56.6**	**76.7**	**62.6**

Bold values indicate our experimental conjecture

Our LEDet-preserving bounding boxes with accurate localization by soft label show significant improvement in AP

of the ship detector. Table 3 compares the detection results of RetinaNet and LEDet on the initial ship datasets without augmentation. It shows that RetinaNet achieves 54.7%AP and LEDet achieves 56.6%AP. Intuitively, we can observe that AP improved 1.9 percentage points (54.7%AP vs 56.6%AP). This confirms the positive effect of the soft label.

As Fig. 8 shows, the detection results of LEDet are more reliable than RetinaNet. The detection score of the bounding box represents the localization quality score of ships, which is a more reasonable representation of the localization quality of ships.

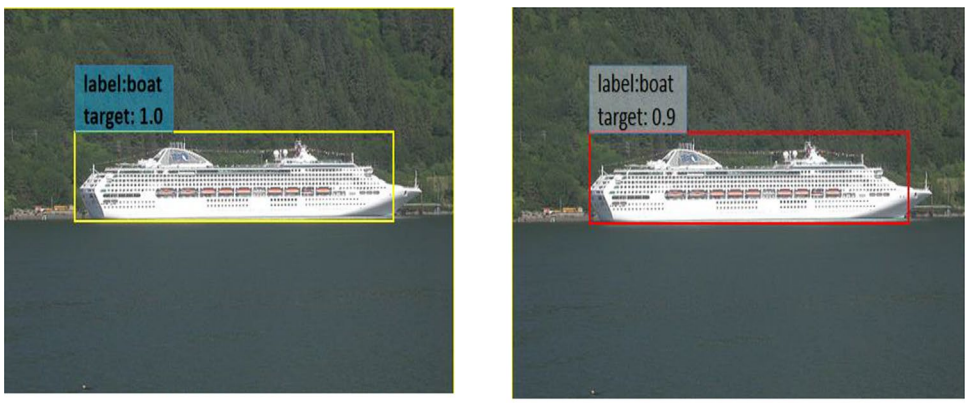
5.2 Datasets augmentation (CTP)

To improve the performance of LEDet, we propose two methods of data augmentation for the initial ship datasets.

The first is to detect the sea line in the new sea pictures, as shown in Fig. 9, and then paste the copied ship image under the sea line to simulate the real scene of the ship, as shown in Fig. 10. The second is to randomly paste the copied ship on the new sea scene even in the sky. For the second method, we consider pasting different numbers of ships on a new sea picture to achieve complicated datasets. We want to augment more kinds of ship datasets by this method, as shown in Fig. 11. Based on the two methods proposed above, we investigate the effect of the two kinds of data augmentations. Table 4 shows our detector with these two data augmentation methods. We find that the accuracy of the ship detection algorithm is higher for the data augmentation with random pasting because the ship positions based on random copying are more diverse. Some pasted ship images even on sky rich the fined features and increase the global feature extraction of the detector. These extraordinary pictures sometimes are similar to flipping the vehicles which do not exist in realistic scenes. Our test datasets consist of all realistic pictures of ships on the sea and the train datasets contain the random pasted ship images even in the sky. Additionally, these augment pictures are only used in the training process. The real pictures are used for validation and testing. Consequently, the main idea is to train an accurate and robust ship detector for USV. Extensive experiments validate the better results of this augmentation method.

5.3 Diverse transformation

To increase the diversity of copied ship targets and randomly paste them into the new sea pictures, we investigate the performance of different transformations for ship targets. Table 5 shows the effects of the operations such as random rotation, random flip, random scaling, block clipping, and random Gaussian noise with 0.5 ratios on the detection performance of LEDet. These results verify the effect of diverse transformations on our LEDet. Table 4 shows that our LEDet with the CTP method achieves 62.2%AP without transformation methods. As shown in Table 5, it is clear that LEDet with random rotation achieves 55.3%AP, about 6.9%AP drop. LEDet with random scaling is only 54.1%AP, about 8.1%AP drop. LEDet with random block clipping is 61.8%AP, about 0.4%AP drop. LEDet with random Gaussian noise is 62.5%AP and the best transformation to our detector is random flip which is further boosted to 62.7%AP. Hence, some discoveries can be found from the effect of diverse transformation. Not all transformations can improve the performance of the detector. Only LEDet with random Gaussian noise (62.5%AP vs 62.2%AP) and random flip (62.7%AP vs 62.2%AP) can improve the performance of LEDet. The remaining methods all decrease the performance of LEDet.



(a) LEDet **(b)** RetinaNet

Fig. 8 **a** RetinaNet uses classification scores to evaluate localization accuracy. **b** LEDet uses soft label localization scores to evaluate localization accuracy. The results show that our LEDet is more reliable for USV to locate ships

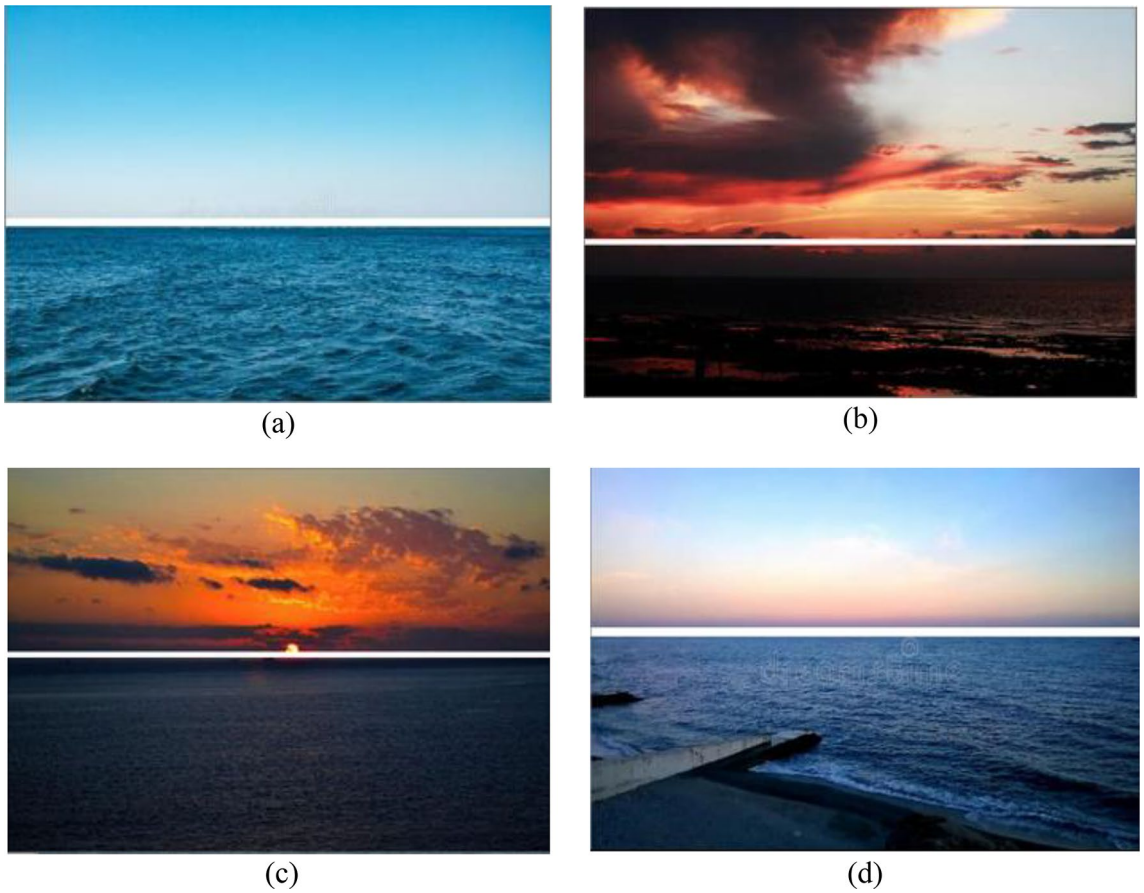


Fig. 9 Sea-sky-line detection of four different sea scene pictures. **a**–**d** are the results

5.4 Comparison with state-of-the-art detector

The LEDet is compared with recent state-of-the-art detectors on our ship datasets, such as RetinaNet, Faster RCNN, YOLOv3, YOLOv4, YOLOv5s, and SSD. Table 6 represents the results. Compare with the powerful baseline RetinaNet, our LEDet achieves 8.0% AP gaps with ResNet-50-FPN (54.7%AP vs 62.7%AP). This validates the contributions of our method. By comparing with the ship detection methods such as YOLOv3. It achieves 9.0%AP gaps with ResNet-50 backbone (53.7%AP vs 62.7% AP). Although recent YOLOv4/YOLOv5s detectors both have better performance than YOLOv3. They all perform worse than our LEDet with random filp. Meanwhile, our one-stage ship detector LEDet is even better than two-stage detector Faster RCNN (62.7%AP vs 56.4%AP). All of the results validate the contribution of our method surpassing almost all recent state-of-the-art ship detectors. Therefore, LEDet has obvious advantages in ship detection and unique excellent characteristics such as self-adaptability, and robustness, which determines its role in future USV perception.

6 Conclusion

To improve the localization accuracy of ship detection on the sea, in this paper, we propose a ship detector LEDet for USV which solves the lack of complete ship datasets and the inaccurate ranking metric of ship detection. The LEDet including CTP and soft labels is proposed to improve the accuracy of ship detection. Based on the ship dataset we generated, the evaluation standard of MS COCO reaches

Fig. 10 Examples of the first data augmentation method to detect the sea line and then paste the ship image under the sea line. **a**–**d** are the results

62.7%AP, which is 8.0%AP higher than RetinaNet and surpasses most of the recent popular object detectors. Therefore, according to experimental results, we conclude that the dataset augmentation method of CTP can increase the complexity and challenge of ship training data. By reorganizing the association and diversity between ship image data and new sea pictures, we can learn the invariant features of the ship and estimate localization accuracy accurately. For soft labels applied in the classification task, ship positions are integrated into the localization score estimation branch as a target to predict a certain category and localization accuracy. This makes the NMS more accurate as it filters the detection results according to the localization accuracy. All of the experiments validate the effectiveness of LEDet, and it can serve as a simple yet effective detector for USVs. Although the experimental study presented in this paper verifies LEDet with data augmentation improving the performance of ship detection. Due to its limitation, some factors are still not satisfactory. Our ship dataset based on the CTP method only increases the similar ship data by transforming the ship images rather than new ship data. Large different ship data is significant for the detector. Besides, the effect of the sea environment such as sunlight, vapor, fog, and so on is sensitive to the performance of LEDet. Further studies are therefore in demand to consider how to augment much more ship data and address the environmental effect on our ship detector LEDet. Hence, the desired ship detector for USV with the developed theoretical method requires investigation in the future.

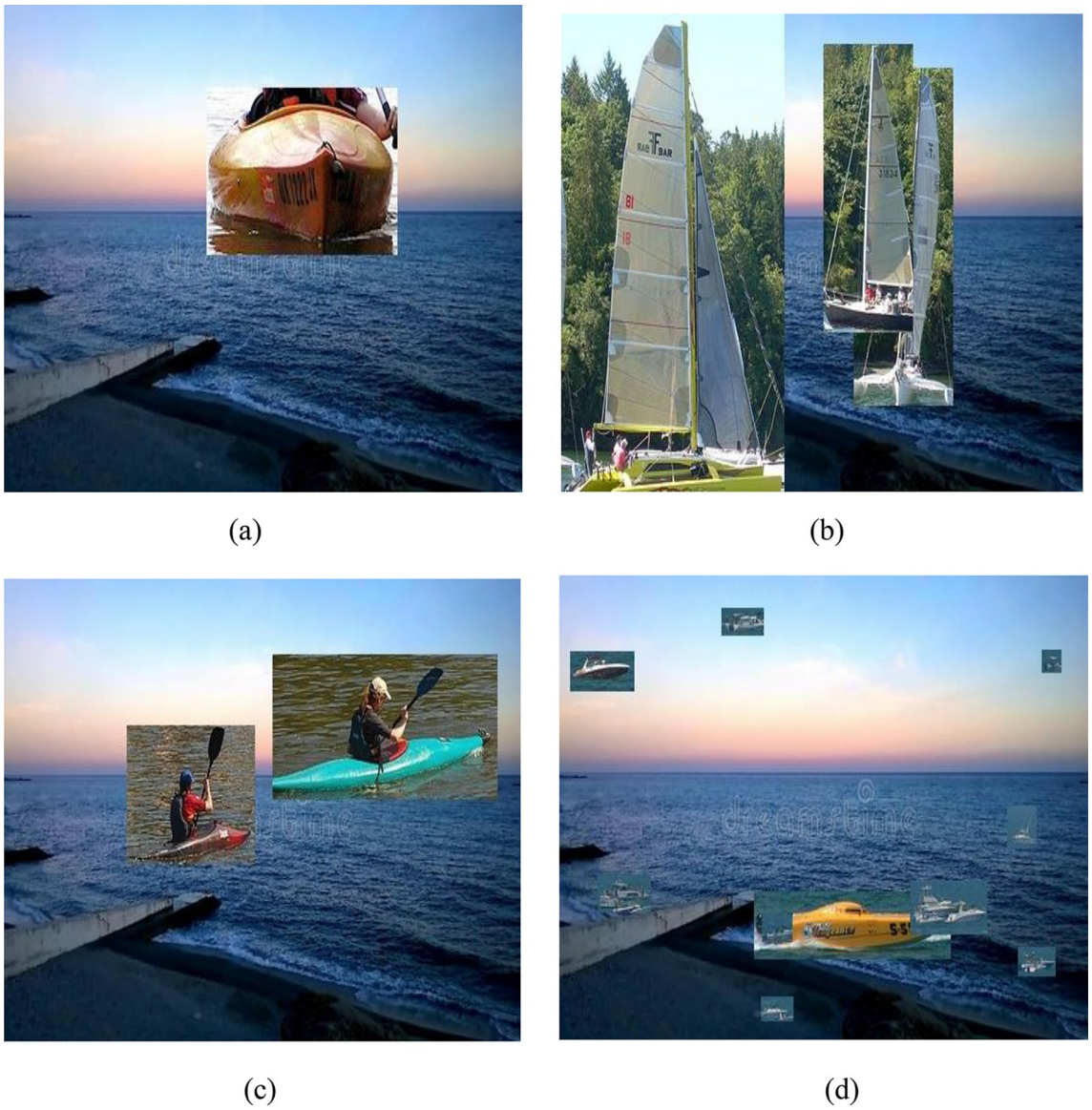


Fig. 11 Examples of the second data augmentation method. To paste the ship images on the new sea pictures randomly. **a**–**d** are the results

Table 4 Comparison of two kinds of data augmentations with LEDet

Method	*AP*	AP_{50}	AP_{75}
LEDet (under sea line)	60.4	82.2	65.5
LEDet (random paste)	**62.2**	**84.3**	**67.8**

Bold values indicate our experimental conjecture

Table 5 Performance of our diverse transformation methods with our proposed LEDet and evaluation on our constructed ship datasets

Method	AP	AP_{50}	AP_{75}
LEDet + random rotation	55.3	79.4	61.6
LEDet + random scaling	54.1	77.7	59.9
LEDet + random gauss noise	62.5	83.9	68.5
LEDet + random block clipping	61.8	84.0	67.6
LEDet + random filp	**62.7**	**84.6**	**68.6**

Bold values indicate our experimental conjecture

Table 6 Performances of recent state-of-the-art ship detectors on our constructed ship datasets

Method	AP	AP_{50}	AP_{75}
RetinaNet	54.7	76.6	60.9
Faster RCNN	56.4	85.0	63.8
YOLOv3	53.7	76.5	60.3
SSD	56.7	82.0	63.7
LEDet + random filp	**62.7**	**84.6**	**68.6**
YOLOv4	57.4	83.4	66.7
YOLOv5s	54.2	76.6	60.4

Bold values indicate our experimental conjecture

Acknowledgements This work was supported by the National Key Research and Development Program of China (No. 2020YFC1521700), Major projects of National Natural Science Foundation of China: 61991415; The Joint Founds of National Natural Science Foundation of China): U1813217; the National Natural Science Foundation of China (No. 51904181). Shanghai Municipal Natural Science Foundation (21ZR1423300).

References

Bochkovskiy, A., Wang, C. Y., Liao, H. Y. M.: Yolov4: Optimal speed and accuracy of object detection. Comput. Sci. (2020). arXiv preprint arXiv:2004.10934

Chen, L., Fukun, B., et al.: An intensity-space domain CFAR method for ship detection in HR SAR images. IEEE Geosci. Remote Sens. Lett. **14**(4), 529–533 (2017). https://doi.org/10.1109/lgrs.2017.2654450

Chen, K., Wang, J., Pang, J., et al.: MMDetection: Open mmlab detection toolbox and benchmark. Comput. Sci. (2019). arXiv preprint arXiv:1906.07155

Chen, Y., Li, Y., Kong, T., et al.: Scale-aware automatic augmentation for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9563–9572 (2021a)

Chen, Z., Ouyang, W., Liu, T., et al.: A shape transformation-based dataset augmentation framework for pedestrian detection. Int. J. Comput. vis. **129**(4), 1121–1138 (2021b)

Cubuk, E.D., et al.: Auto augment: Learning augmentation strategies from data. IEEE/CVF Conf. Comput. vis. Pattern Recognit. (CVPR) (2019). https://doi.org/10.1109/cvpr.2019.00020

Cubuk, E.D., Zoph, B., Shlens, J., et al.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)

Duan, K., Bai, S., Xie, L., et al.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6569–6578 (2019)

Fefilatyev, S., Goldgof, D., Shreve, M., et al.: Detection and tracking of ships in open sea with rapidly moving buoy-mounted camera system. Ocean Eng. **54**, 1–12 (2012). https://doi.org/10.1016/j.oceaneng.2012.06.028

Fingas, M.F., Brown, C.E.: Review of ship detection from airborne platforms. Can. J. Remote. Sens. (2014). https://doi.org/10.1080/07038992.2001.10854880

Girshick, R.: Fast R-CNN. Comput. Sci. (2015). https://doi.org/10.1109/iccv.2015.169

Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Comput. Soc. (2013). https://doi.org/10.1109/cvpr.2014.81

He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. IEEE Conf. Comput. vis. Pattern Recognit. (CVPR) (2016). https://doi.org/10.1109/cvpr.2016.90

He, K., Gkioxari, G., Dollár, P., et al.: Mask R-CNN. IEEE (2017). https://doi.org/10.1109/iccv.2017.322

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. IEEE/CVF Conf. Comput. vis. Pattern Recognit. (CVPR) **2020**, 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975

Jia, D., Wei, D., Socher, R., et al.: ImageNet: A large-scale hierarchical image database. Proc. of IEEE Comput. vis. Pattern Recognit. (2009). https://doi.org/10.1109/cvprw.2009.5206848

Jiang, B., Luo, R., Mao, J., et al.: Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV), pp. 784–799. Springer, Cham (2018)

Kong, T., Sun, F., Liu, H., et al.: Foveabox: Beyound anchor-based object detection. IEEE Trans. Image Process. **29**, 7389–7398 (2020)

Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. (2012). https://doi.org/10.1145/3065386

Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)

Lecun, Y., Bottou, L.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791

Li, X., Wang, W., Wu, L., et al.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Adv. Neural. Inf. Process. Syst. **33**, 21002–21012 (2020)

Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft COCO: common objects in context. In: European conference on computer vision. Springer International Publishing, Berlin (2014). https://doi.org/10.1007/978-3-319-10602-1_48

Lin, T.Y., Dollar, P., Girshick, R., et al.: Feature pyramid networks for object detection. IEEE Conf. Comput. vis. Pattern Recognit. (CVPR) (2017). https://doi.org/10.1109/cvpr.2017.106

Neubeck, A., Gool, L.: Efficient non-maximum suppression. Int. Conf. Pattern Recognit. (2006). https://doi.org/10.1109/icpr.2006.479

Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. IEEE (2017). https://doi.org/10.1109/cvpr.2017.690

Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. comput. Sci. (2018). arXiv preprint arXiv:1804.02767

Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031

Rezatofighi, H., Tsoi, N., Gwak, J.Y., et al.: Generalized intersection over union: a metric and a loss for bounding box regression. IEEE/CVF Conf. Comput. vis. Pattern Recognit. (CVPR) (2019). https://doi.org/10.1109/cvpr.2019.00075

Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. comput. Sci. (2014). arXiv preprint arXiv:1409.1556

Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)

Tang, Y., Li, B., Liu, M., et al.: Autopedestrian: an automatic data augmentation and loss function search scheme for pedestrian detection. IEEE Trans. Image Process. **30**, 8483–8496 (2021)

Tian, Z., Shen, C., Chen, H., et al.: FCOS: Fully convolutional one-stage object detection. IEEE/CVF Int. Conf. Comput. vis. (ICCV) (2019). https://doi.org/10.1109/iccv.2019.00972
Wu, S., Li, X., Wang, X.: IoU-aware single-stage object detector for accurate localization. Image vis. Comput. **97**, 103911 (2020)
Wu, S., Yang, J., Wang, X., et al.: Iou-balanced loss functions for single-stage object detection. Pattern Recognit. Lett. **156**, 96–103 (2022)
Zhang, Y., Li, Q.Z., Zang, F.N.: Ship detection for visual maritime surveillance from non-stationary platforms. Ocean Eng. **141**, 53–63 (2017)
Zhi, Z., Ji, K., Xing, X., et al.: Ship surveillance by integration of space-borne SAR and AIS–review of current research. J. Navig. **67**(1), 177–189 (2014). https://doi.org/10.1017/s0373463313000659
Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 850–859 (2019)
Zhu, C., Chen, F., Shen, Z., et al.: Soft anchor-point object detection. In: European conference on computer vision, pp. 91–107. Springer, Cham (2020)
Zoph, B., Cubuk, E.D., Ghiasi, G., et al.: Learning data augmentation strategies for object detection. In: European conference on computer vision, pp. 566–583. Springer, Cham (2020)

Dr.Yang Zhou received his undergraduate Bachelor's degree in Mechanical Engineering from Southwest Petroleum University (Chengdu, China) in 2013. He then completed his Ph.D in Fluid dynamics from Fudan University (Shanghai, China) in 2018. During 2015 to 2017, he acted as a visiting student in Federal University of Rio de Janeiro (Rio de Janeiro, Brazil) and China University of Petroleum, Beijing (Beijing, China). Currently, he is a research associate at Research Institute of USV Engineering in Shanghai University (Shanghai, China). His current research aims at understanding the representation of computer vision and image processing with theoretical description. His research interests also focus on Artificial Intelligence & Intelligent unmanned system.

Jingling Lv received the B.E. degree in Department of Mechanical engineering, Anhui Agricultural University (Anhui, China), in 2019. He is now a master course student of Shanghai University (Shanghai, China). His research interests include computer vision and image processing.

Yueying Wang (M'16-SM'18) received the B.Sc. degree in mechanical engineering and automation form the Beijing Institute of Technology, Beijing, China, in 2006, the M. Sc. degree in navigation, guidance, and control, and Ph.D. degree in control science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2010 and 2015, respectively. He is currently a Full Professor with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai. His current research interests include intelligent and hybrid control systems, control of unmanned aerial/surface vehicles. He has served on the editorial board of a number of journals, including IET-Electronics Letters, International Journal of Electronics, International Journal of Fuzzy Systems, International Journal of Control, Automation and Systems, Journal of Electrical Engineering & Technology, and Cyber-Physical Systems.

Chang Liu is currently a Ph.D. student of Unmanned Surface Vehicle Engineering Research Institute in Shanghai University. His research interests include pattern recognition, computer vision, and video analysis.

Songyi Zhong received the B.Sc. and the Ph.D. in mechanical engineering from Northwestern Polytechnical University, China. Currently, he is an assistant professor with Shanghai University, China.

Guozhu Tan is currently a master student at the School of Computer Science and Engineering in Shanghai University. His research interests include computer vision and text recognition.

Jiacheng Sun is currently a M.D. student of School of Computer Enginerering and Science in Shanghai University. His research interests include wireframe parsing and linear object detection.