IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 14, 2021, accepted December 16, 2021, date of publication December 27, 2021,
date of current version January 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3138983

Water Target Recognition Method and Application for Unmanned Surface Vessels

LIANG CHENG[1,2], BAOYUAN DENG[3], YUAN YANG[3], JIFANG LYU[1,2], JICHENG ZHAO[2], KE ZHOU[3], CHUNLI YANG[2], LEIGANG WANG[2], SHIYUAN YANG[2], AND YUNZE HE[3]

[1]School of Ocean Engineering, Jiangsu Ocean University, Lianyungang 222005, China
[2]Zhuhai Yunzhou Intelligent Technology Company Ltd., Zhuhai 519085, China
[3]College of Electrical and Information Engineering, Hunan University, Changsha 410006, China

Corresponding author: Yuan Yang (yangyuan96work@163.com)

This work was supported in part by Zhuhai Yunzhou Intelligent Technology Company Ltd., under Grant YZ-LX0A1-820.

ABSTRACT Water target recognition is a critical challenge for the perception technology of unmanned surface vessels (USVs). In the application of USV, detection accuracy and the inference time both matter, while it is tough to strike a balance and single-frame water target detection behaves unstable in the video detection. To solve these problems, many strategies are applied to increase YOLOv4's performance, including network pruning, the focal loss function, blank label training, and preprocessing with histogram normalization. The optimized detection method achieves a mean average precision (mAP) of 81.74% and a prediction speed of 26.77 frames per second (FPS), which meets the USV navigation requirements. To build the integrated USV-based system for water target recognition, a water target dataset containing 9936 images is created from offshore USV experiments in which the human-in-the-loop annotation and mosaic data augmentation methods are used. The issues of miss detection and false alarm can be considerably mitigated by cascading the Siamese-RPN tracking network, and the major color of a water target can be retrieved using a local contrast saliency color detection scheme. The system being tested is called ''ME120'' includes an embedded edge computing platform (Nvidia Jetson AGX Xavier). Finally, online dataset learning demonstrates the improved YOLOv4 achieves an increase of 66.98% in FPS at the cost of a decrease of 0.79% in mAP when compared with the original YOLOv4 and offline navigation experiments validate that our system achieves high recognition capability while maintaining a high degree of robustness.

INDEX TERMS Object recognition, USV, deep learning, computer vision.

I. INTRODUCTION

USVs will be critical in the future water surface environment. It is useful in a variety of industries, including marine emergency rescue, logistics, water quality monitoring, hydrological surveying, marine environment mapping, and water ecological protection. In comparison to traditional manual vessels, the USV can work ceaselessly in areas that are inaccessible to humans, ensuring consistency in terms of execution efficiency and judgment quality. With the rapid advancement of computer vision technology, it is becoming easier for the USV outfitted with a visual intelligent perception system to efficiently recognize water targets. As a result, our research on the method and application of water target recognition is both theoretically and practically significant.

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

USVs' conventional sea surface perception technology is based on millimeter-wave radar, lidar, inertial measurement units, and GPS, among other sensors. Perception techniques based on computer vision have advanced dramatically in recent years. Optical pictures contain more comprehensive information in the region of interest (ROI) of the target area, making it easier to discriminate the water target using visual perception techniques. Specifically, using the video stream acquired by onboard visual sensors in conjunction with sophisticated object detection technology, it is possible to achieve exceedingly accurate and rapid object recognition and localization.

This study examines an artificial intelligence system based on a vision sensor and an edge computing platform for recognizing and tracking a water target that must meet the accuracy and real-time requirements. As illustrated in Figure 1, the ship is a representative of Zhuhai Yunzhou Intelligent

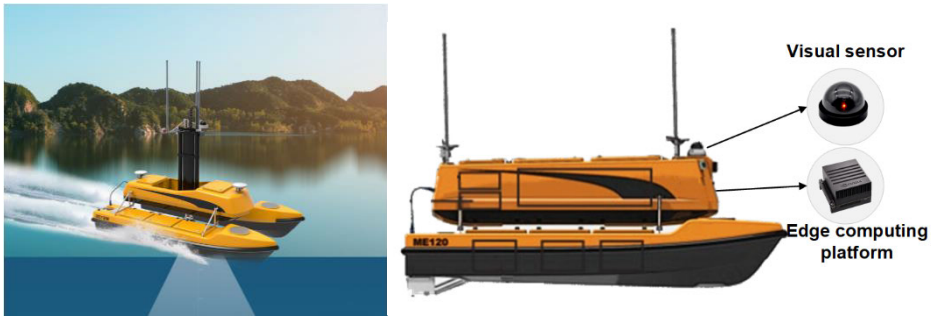

FIGURE 1. The ME120 USV with visual sensor and edge computing platform. The edge computing platform refers to NVIDIA Jetson AGX Xavier Developer Kit (or similar edge computing platform). Which supports NVIDIA Jetpack, DeepStream SDK, and other Deep-Learning software packs, including CUDA, cuDNN, TensorRT, etc.

Technology Co., Ltd.'s environmental monitoring USV. A high-definition photoelectric vision sensor can be included into the ship's completely autonomous lifting fin system. This USV is capable of intelligent obstacle avoidance via real-time transmission of high-definition water surface footage. USV's perception technology is primarily used to perform accurate obstacle avoidance and shore docking.

As a result, the USV's capacity to detect surface objects is necessary to resolve this issue. However, there are certain difficulties in detecting water targets: 1) It is difficult to strike a balance between detection precision and inference time for the detection of actual water targets. For example, recognition accuracy (confidence level 50% or greater) within a 10 meter distance must be at least 80% or greater for real-time ship operation. Simultaneously, the algorithm can only access the embedded platform, whereas the object detection algorithm's inference speed must exceed 25 frames per second to meet the requirement of real-time operation and to match the speed of the video stream; 2) In addition to single-frame target detection, other methods are required in actual application of video object detection. The single-frame target detection method is susceptible to problems such as tiny objects, poor illumination, and imbalanced data.

A water target recognition system based on USV intelligent perception is proposed to address these issues. Our contributions are summarized as follows: 1) A ''improved'' YOLOv4 is proposed with comparable mAP and significantly increased FPS. Several methods, including network pruning, focal loss function training, blank label training, and histogram normalization preprocessing, are used to improve the performance of YOLOv4; 2) A system was established by combining dataset configuration, training tricks, and a three-stage scheme for water target detection, tracking, and color detection. This system has been validated in real-world environments.

The remainder of this paper is organized as follows. Section II summarizes new research in the field and compares prominent object detection techniques. Section III creates water target datasets using images from the internet and images obtained by USV. Section IV enhances the YOLOv4 algorithm to meet the requirements of USV navigation. By including the Siamese-RPN tracking algorithm and the LC Saliency color detection technique, the water target system gains a greater capacity for information perception, as detailed in Section V. The method of deployment and the results of the experiments are described in Section VI. Section VII concludes the paper.

II. RELATED WORK

The field of water target recognition and auxiliary human eye observation has benefited greatly from image and video processing technologies. Strickland and Hahn [1] extracted the ship's outline from an infrared image using the wavelet transform filtering approach. Withagen *et al.* [2] classified the ship using the forward-looking infrared camera image. Smith and Teal *et al.* [3] investigated the visual image system used to recognize high-speed passenger ferries. Wen-Jing *et al.* [4] enhanced the Mean-Shift segmentation approach for extracting the water target and tracked the target's position using the Kalman filter. Shi *et al.* [5] recognized the ship target by using the frequency-tune saliency extraction algorithm and followed it using the correlation filter approach for joint location and scale estimation. However, the correlation filtering method makes real-time tracking in occlusion challenging. Generally, early visual perception technology is constrained by its era, relying heavily on wavelet transforms, manual feature extraction, and wave filters, among other techniques, making it difficult to complete identification and tracking of complex targets.

After entering the era of deep learning, the great feature extraction capability of convolutional neural network enables it to perform well in the field of water target recognition. Peng *et al.* [6] recognized six distinct types of water objects using the Faster R-CNN approach, which achieved an AP of 91.97% in the datasets of 1189 images. Li *et al.* [7] proposed a multiscale object detection CNN based on an enhanced Faster R-CNN that detects ships in optical remote sensing images with high efficiency. Wei [8] implemented automatic recognition and tracking on the ''WAM-V-USV'' platform, which uses the YOLOv3 algorithm. However, given the constraints imposed by objective conditions, there is still potential for improvement in the technique for extracting

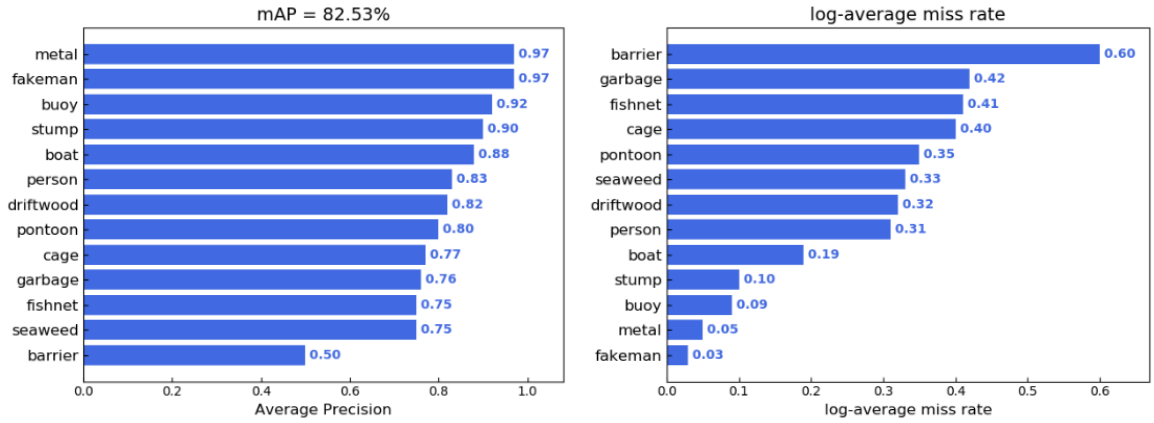

FIGURE 2. **The detection results of YOLOv4. Most of the category targets can be accurately detected, only the barrier has a high false recognition rate.**

neural network features. Bochkovskiy *et al.* [9] presented a new real-time object detection system, YOLOv4, in the year 2020, based on numerous unique detection approaches developed over the previous two years. It attained the highest mean average precision (mAP) and inference speed on MS COCO datasets as a new effort in the YOLO series. In conclusion, the deep learning method can achieve the level of accuracy that older methods cannot in this industry. The introduction and advancement of new deep learning-based visual object identification technology will be a prominent development trend in the future.

TABLE 1. **Comparison of networks performance.**

Algorithm	mAP1 (%)	mAP2 (%)	Speed1 (FPS)	Speed2 (FPS)
Faster R-CNN [10]	45.21	/	3.82	/
SSD [11]	31.18	/	19.74	/
YOLOv3 [12]	67.06	48.87	22.35	14.41
YOLOv4 [9]	82.53	80.12	41.43	15.01

We tested four common deep neural networks on water target datasets in this article, and the detection results are listed in Table 1. We evaluate network training performance primarily in terms of inference speed and recognition accuracy. The FPS metric is used to determine the inference speed, and the mAP metric is used to determine the recognition accuracy. The mAP demands a better than 50% intersection over union (IoU) ratio between the detected bounding box and the ground truth bounding box, which is determined using the Pascal VOC Challenge formula.

The mAP1 and Speed1 columns in Table 1 correspond to the findings obtained on a high-performance machine equipped with an Intel i9-7920x CPU and an Nvidia RTX2080ti GPU. While the mAP2 and Speed2 are the outcomes of the NVIDIA Jetson AGX Xavier real-time deployment platform. Due to cost considerations, Faster R-CNN and SSD were not deployed, leaving the related data in Table 1 empty. Due to the fact that the computational performance of an edge computing platform is significantly less than that of a high-performance computer, the average precision of the identical method may be reduced upon deployment. When the two indices are compared, YOLOv4 has the highest average precision, which will serve as the foundation for our detection system. However, its inference speed is just 15.01 frames per second, which falls short of the requirements for onboard real-time detection. It will be enhanced in Section IV to take into account the features of the target datasets and the requirements for deployment to an edge computing platform. Fig. 2 illustrates the specific detection accuracy of YOLOv4, in which most of the category targets can be accurately detected. Only the barrier has a high false recognition rate.

III. THE CONSTRUCTION OF WATER TARGET DATASETS

The quantity and quality of the datasets is a significant influencing factor on the object detection method based on deep learning strategy.

The water target primarily considers two application scenarios: the first is object detection for obstacle avoidance, which includes stumps, driftwood, seaweed, buoys, garbage, people, and boats, among others; and the second is object detection for operation functions, which includes fishnets, barriers, cages, fakemen, and pontoons, among others. Several common targets are displayed in Figure. 3, which may represent the primary objectives of USV's artificial intelligence system.

There are four primary methods for obtaining water target images: via open-source datasets, via mobile phone photographs, via USV acquisition, and via web crawler. A eligible image that fits the requirements undergoes numerous phases during data processing, including frame extraction, duplication removal, manual filtering, label annotation, and data augmentation. Because the long-term movies acquired with USV perception technology typically contain relatively

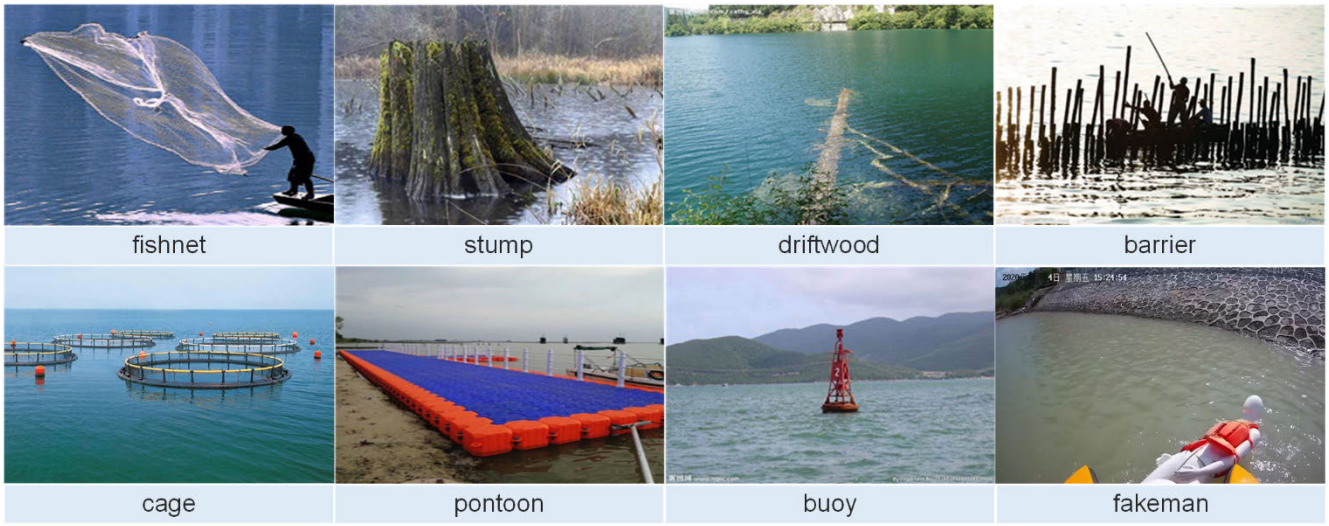

FIGURE 3. Examples of water targets.

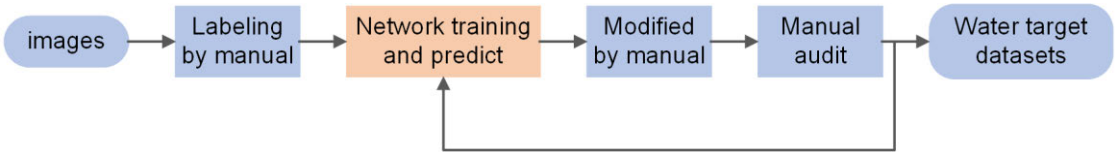

FIGURE 4. The method of annotating using a human-in-the-loop. To begin, we need manually label a set of images and train the network synchronously until the network demonstrates preliminary detection capacity. New images will be predicted using the network method. Then, based on the network prediction results, the professions manually adjust and audit. Continue this process until we have all the labeled images and a network algorithm that works well.

similar scenes and objects, some frame extraction processing, such as collecting a picture every 25 frames, is required. Then, using the structural similarity (SSIM) [13] algorithm, contrast analysis is utilized to eliminate duplications and images that have more than 80% similarity. Following that, manual screening will exclude undesirable photographs, such as those with low quality and missing targets, as well as ones that are too difficult to recognize.

A. HUMAN-IN-THE-LOOP ANNOTATION METHOD

The next phase is label annotation, which is a time-consuming and tedious process. It is difficult for a manual to do labeling work for an extended period of time, adhering to tight standards and producing high-quality work. However, the human-in-the-loop method [14] can significantly reduce workload. Fig. 4 illustrates the process of intelligent data labeling.

For instance, the process of a human-in-the-Loop data annotation is logged in Table 2. Our team contributed approximately eight members to this effort. Initially, we employed the standard labeling approach, which was extremely slow, costing approximately 50 days for 2913 effective labels. In the subsequent stage, 4127 labels were labeled in around 20 days, resulting in a 3.5-fold increase in labeling efficiency.

TABLE 2. An instance of human-in-the-loop annotation.

Datasets	Labels	Cumulative	Strategy	Time of consuming
I	1717	1717	Traditional manual working	30 days
II	1196	2931	Traditional manual working	20 days
III	1739	5913	Human-in-the-loop	10 days
IV	2388	8301	Human-in-the-loop	10 days

While this is an approximation, the efficiency improvement associated with the human-in-the-loop method is obvious.

B. MOSAIC DATA AUGMENTATION METHOD

To compensate for the lack of datasets, it is usually necessary to perform some data augmentation. Typical data augmentation techniques include horizontal/vertical flipping, rotation/reflection, random scaling, random cropping, color jittering, and adding noise. Additionally, Glen Jocher [9], [12] advocated mosaic data augmentation as a strategy of more efficiently enhancing data. In contrast to conventional

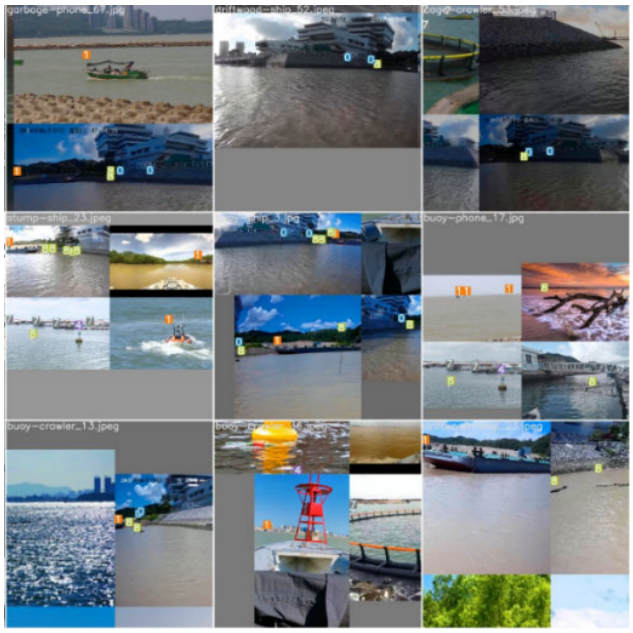

FIGURE 5. A new method of data augmentation: Mosaic. During training time, the network will randomly extract four images, and then extract a random point near the center point of the augmented image, and then align the extracted four images with this vertex through scaling and clipping, and finally obtain a new image.

FIGURE 6. Examples of training samples.

TABLE 3. Water target datasets composition.

Class	The number of images	The number of labels
Fishnet	957	1023
Stump	1531	2596
Barrier	1834	2155
Cage	272	692
Person	2172	4390
Driftwood	1106	1897
Seaweed	183	281
Pontoon	356	406
Buoy	703	971
Garbage	842	1405
Boat	5566	9620
Fakeman	976	1111
Metal	3568	21235

image data augmentation approaches [15], the mosaic data augmentation method splices numerous images (often four) into a single complete image, which has been shown to be more successful for training.

The mosaic data preprocessing method is used in this article, as illustrated in Fig. 5. After mosaic augmentation, the region of large targets shrinks, allowing for increased training of small targets, which benefits network recognition of small targets.

C. WATER TARGET DATASETS

Following frame extraction, duplication removal, manual filtering, human-in-the-loop label annotation, and data augmentation, 9936 photos are chosen as the final water target datasets. The dataset contains 13 different types of objects categorized as fishnet, stump, barrier, cage, human, driftwood, seaweed, pontoon, buoy, garage, boat, fakeman, and metal. Table 3 summarizes the composition of the water target dataset.

Due to the fact that a single image may contain numerous categories and targets, the total number of photos is the sum of all images that contain at least one of these targets. By and large, these multi-target detection datasets suffer from a serious sample imbalance problem.

Fig. 6 illustrates several targets. As can be observed, the background region of photographs changes dramatically, as does the label size of the target. For instance, some distant boats have a pixel size of 10×10 pixels. While the barrier in the foreground has a wide pixel area, it can take up more than 60% of the entire image.

IV. IMPROVEMENT OF OBJECT DETECTION ALGORITHM

There are two issues with implementing YOLOv4: 1) the inference speed of operation on the edge computing platform is suboptimal; 2) the phenomenon of incorrect recognition for a certain class is readily apparent. The following are the corresponding improvement measures.

A. NETWORK PRUNING

Although the YOLOv4 network's backbone "CSPDarknet-53" offers exceptional feature extraction capability, the repeating of gradient weights during network optimization results in a huge amount of computation and a high inference

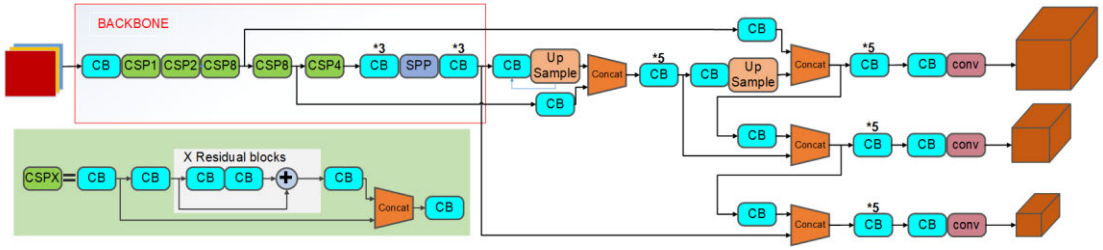

FIGURE 7. The structure of pruned YOLOv4. The "conv" module represents the convolution layer. The "CB" represents the structure contains a conv layer and a Batch Normalization [16] layer. The "Concat" means the combination or a shortcut of two feature maps. The "Up Sample" represents an up-sampling operation, which makes the feature maps greater. And the "CSPX" module represents CSPDarknet structure, which combines the Cross Stage Partial Network [17] and ResNet [18], and the "X" means how many residual components are used. The "SPP" module represents the spatial pyramid pooling method [19]. All module structures can be seen in paper [9].

TABLE 4. Comparison of each enhancement algorithm.

method	mAP(%)
Origin	79.25
Linear transformation	78.96
Histogram normalization	80.03
Gamma transformation	78.31
Global histogram equalization	77.35
CLAHE	78.63

cost, making it unsuitable for deployment to edge computing platforms. As a result, the backbone is appropriately lowered in size to increase the network algorithm's inference speed at the expense of somewhat reduced average precision.

Fig. 7 illustrates the network structure after pruning. The core concept of YOLOv4 stays the same, in which the input module, neck, and prediction module have been lowered in size. Only the backbone module has been decreased in size. The backbone module in its original form contains 73 convolution layers. After pruning, 60 convolution layers are used, significantly reducing computation.

Section IV.E Ablation experiment illustrates the performance of the pruned YOLOv4 on the water target datasets. Although 3.28% of mean Average Precision is sacrificed in comparison to the original, the inference speed is significantly increased by 66.81%, meeting the basic requirements for USV detection.

B. FOCAL LOSS FUNCTION

Focal loss [20] is a loss function method for resolving the object detection problem of imbalanced ratios of positive and negative samples. The loss function, as the underlying principle of difficult sample mining, reduces the weights of several negative samples during the network optimization process.

$$FL(p) = -\alpha(1-p)^{\gamma}\log(p) \tag{1}$$

As shown above, p represents the prediction probability, α and γ are the preset adjustment parameter.

In a nutshell, focused loss is the process of augmenting the cross-entropy loss function with adjustment parameters. α and γ are concentration parameters that can be adjusted to change the influence of simple samples seamlessly. Table 5 Ablation experiment shows the experimental outcomes from this paper after applying the focal loss function to the pruned YOLOv4. As can be seen, the difficult-to-identify barrier class has been improved, while the mAP has also been marginally improved.

C. BLANK LABELS TRAINING

In contrast to human cognition, when a neural network encounters "new" items or scenes (such as categories not included in the training set), it will make significant recognition errors, which is referred to as a "fooling image" [21]. In practice, this manifests as severe false positives, when the background or unknown object is recognized as some type of water target in the datasets.

As illustrated in Fig. 8(a), the CNN model previously displayed a significant misidentification phenomenon during actual USV navigation. Due to the vision sensor's location onboard, it's impossible to avoid shooting the surrounding support pole and antenna cable. However, the water target detection technology consistently incorrectly detects those as navigation buoys located far from the sea.

To address this grave issue, it is vital to manage the network's false recognition rate in practice. As a result, images with blank labels are used for network training, serving as negative samples. By lowering the confidence in unknown class recognition, the negative sample lowers the false positive rate. Figure 8(b) depicts the experimental outcome. Since that time, there have been almost no instances of false identification.

It should be noted that this strategy will also improve recognition accuracy; there is a trend toward increasing mAP as the number of blank labels increases. As demonstrated in Table 5 Ablation experiment, the mean Average Precision of the detection findings increases to 81.27 % after adding the blank label.

However, given the robustness and generalizability of the network approach, this paper only adds 10% blank labels

TABLE 5. Ablation experiment.

	mAP1(%)	mAP2(%)	Speed1(FPS)	Speed2(FPS)
Origin	82.53	80.12	41.40	15.01
Network pruning	79.25	77.29	69.06	26.75
Network pruning + Focal Loss	79.58	77.68	69.03	26.72
Network pruning + Blank labels training	81.27	79.20	69.05	26.75
Network pruning + Histogram Normalization	80.03	78.34	69.01	26.65
Network pruning + Focal Loss + Blank labels training + Histogram Normalization	81.74	79.30	69.13	26.77

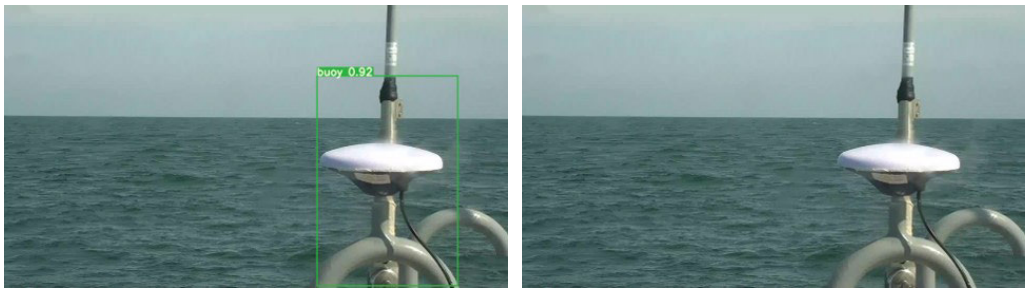

FIGURE 8. Comparison before and after the application of blank labels training.

(for example, if there are 7452 photos in the train datasets, 745 images with blank labels are added), while improving performance (mAP index) by 2.02%. With the addition of more blank labels, the gain in detection accuracy may be more noticeable. However, because this strategy may interfere with the generalization of the network algorithm, it should be used sparingly in practice.

D. PREPROCESSING BASED ON SPECIAL WEATHER

In different weather conditions, such as rain, fog, night, and so on, the detection performance of the algorithm can be improved after image enhancement. In the case of bad weather conditions, the gray value of the image will decrease, resulting in some targets cannot be recognized by the network. References [9], [12] mentioned that data enhancement of images before input into the network can effectively improve detection accuracy.

The common methods of image enhancement are summarized as follows: 1) The whole or local characteristics of the image can be emphasized to make the original image clear; 2) It is suggested that the differences between the features of different objects in the image can be enlarged by emphasizing some interesting features; 3) The suppression of the features of interest can improve the image quality, enrich the information quantity, and enhance the image interpretation and recognition effect.

In this paper, five popular image enhancement methods are applied, including linear transformation [22], histogram normalization [23], gamma transform [24], global histogram equalization [25], and contrast limited adaptive histogram equalization (CLAHE) [26]. The experiment's concept is as follows: the water target image is enhanced and then detected using the pruned network algorithm, and the results are compared to the original image without preprocessing, in order to investigate the effect of image enhancement on water target recognition in various environments. It will be determined during the enhancement operation by the image's gray value. Generally, the sample image's average grayscale value will be calculated and compared to the threshold value (for example, 120). When the image's average grayscale value is larger than 120, no image preprocessing is required. On the other hand, if the threshold value is less than 120, the image is not sufficiently clear and should be enhanced.

The effect comparison of each enhancement algorithm is shown in Table 4. The original detection's mean Average Precision can reach 79.25% without image enhancement. Other algorithms have a slight negative impact on the network detection, while histogram normalization has a certain improvement on the performance, which can be used as a practical enhancement method.

Generally, image enhancement should be used with caution during adverse weather conditions such as gloomy, rainy, or foggy days. Specific examination of a particular problem, if performed harshly on large datasets, may be counterproductive.

In this paper, we take histogram normalization method for image enhancement on rainy and foggy days. Fig. 9 shows the comparison before and after histogram normalization. Table 4 shows that the detection accuracy after

FIGURE 9. Comparison of before and after histogram normalization.

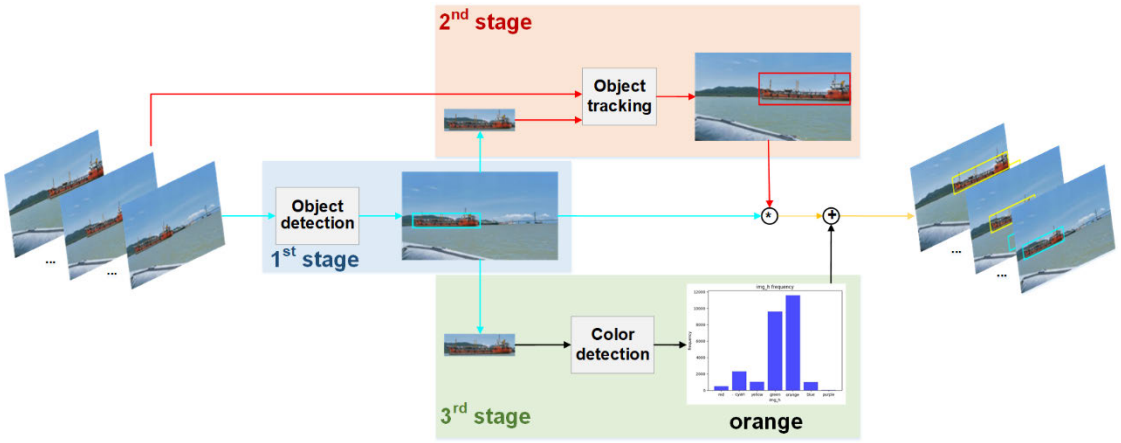

FIGURE 10. The water target detection-tracking-color three-stage scheme.

histogram normalization reaches 80.03%. Compared with the original detection, the mAP value increased by 0.78%. Especially, only the average precision value of boat class decreased slightly, while the AP value of all other classes increased.

E. ABLATION EXPERIMENT

A comparison ablation experiment was conducted using the above-mentioned enhanced approaches. Table 5 summarizes the experimental findings. The mAP1 and Speed1 metrics correspond to findings obtained on a high-performance computer equipped with an Intel i9-7920x CPU and an Nvidia RTX2080ti GPU. While the mAP2 and Speed2 are the outcomes from the NVIDIA Jetson AGX Xavier real-time deployment platform.

Since the inference speed of the original algorithm after deployment is only 15 FPS, which is far from the basic requirements of USV real-time detection. After network pruning, the network's inference speed reached 26 FPS, which fully meets the requirements of USV, while on the cost of slightly detection accuracy decent.

V. OBJECT TRACKING AND COLOR DETECTION ALGORITHM

Based on the optimization methods above, the USV intelligent sensing system has satisfied the requirements of single image detection preliminarily. However, object detection in the continuous image sequences (video) and additional information about the object are needed for the navigation of USV. Therefore, this section mainly studies the application of water target recognition based on the improved YOLOv4.

In order to meet the actual requirements of the automatic navigation of USV, the water target recognition-tracking-color three-stage scheme is proposed in this paper. As shown in Fig. 10, the whole scheme is based on the object detection network (the improved YOLOv4 algorithm above). When the system starts to read the input video stream, the water target detection network will output the category, confidence, and bounding box of the object of interest in the first stage. In the second stage, the object tracking network will use the result of the object detection network as the tracking template. Ship targets are usually set as the tracking objects, and the results of the tracking network and the detection network will be fused to generate the final output in the object detection-tracking fusion scheme. In the third stage, the color recognition algorithm will take the cropped sub-image according to the detected bounding box as the input image, and extract the main color of the target based on the saliency recognition of the foreground. Navigation buoys and ship targets are usually set as color detectable objects. The details of object tracking and color detection algorithms will be introduced in this section.

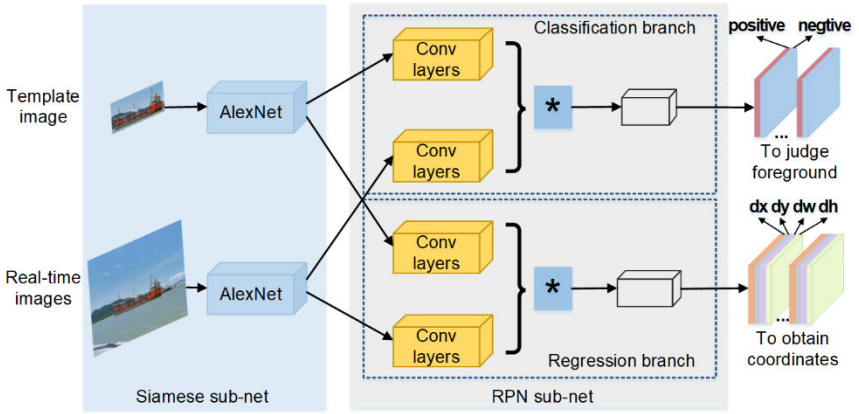

FIGURE 11. The structure of the Siamese-RPN tracking network algorithm.

A. CASCADE OBJECT TRACKING METHOD BASED ON SIAMESE-RPN NETWORK

If the object detection network is fed with frames in a video frame-by-frame, the temporal context is missed, thus causing unstable detection results, and even miss recognition. Because there is a lot of redundancy between adjacent frames, it is feasible to speed up the video detection without hurting performance or to enhance the stability of video detection [27]. To improve its stability in the actual navigation of USV, this section studies how to use the object tracking algorithm to assist the object detection algorithm in a cascade way. The improved YOLOv4 also encounters the same problem, when it is used for frame-by-frame video detection. The shortcomings are concluded as follows:

1) In terms of accuracy, the object detection network with high mAP suffers missed detection in long-time frame-by-frame video recognition;

2) In terms of inference performance, the speed of frame-by-frame object detection may be unstable.

To cover these two issues, a special target tracking network is introduced as compensation. The object tracking network algorithm will work by taking the template obtained from the object detection network.

Object tracking is the process of locating moving objects in video over time. According to the number of tracked targets, it is classified as single object tracking (SOT) and multi-object tracking. In this paper, we focus on using SOT to enhance the stability of video detection. Early approaches in SOT try histogram of oriented gradient (HOG) features and kernelized correlation filters (KCF) to get the heatmaps of the template image and the real-time image. Then, the HOG features are replaced by a fully convolutional network for the feature extraction of both the template image and the real-time image [28]. Due to the same network structure and weights, this network is called the Siamese fully convolutional (Siamese-FC) network. Later, the Siamese-RPN network algorithm [29] is proposed by cascading a region proposal network [10] after the Siamese-FC. In this paper, we use the Siamese-RPN to track targets such as ships.

As shown in Fig. 11, the network structure, this model is composed of a Siamese sub-network for feature extraction and an RPN sub-network for the determination of template coordinates. In the Siamese sub-network, a simple improved AlexNet [30] is used to generate the feature maps. In the RPN sub-network, it is divided into two branches, in which the classification branch determines background or foreground, and the regression branch regresses the bounding box coordinates. In detail, the KCF algorithm is applied to the feature map of the template image and the feature map of the real-time image first, resulting in heatmaps in feature map level. The heatmaps will be used to judge the foreground and regress the coordinate. Generally, this algorithm has achieved accurate results with real-time inference speed, however, Siamese-RPN also suffers when the image changes quickly.

In order to improve the stability of video detection, an object detection-tracking fusion scheme is proposed in this paper. As shown in Fig. 12, the model is initialized (the confidence threshold C_0 and maximum tracking frames T_0 are set) at the beginning. Then the improved YOLOv4 algorithm will detect the input video frame and output the bounding box D_1, confidence C_1 and class L. After outputting the first frame, it will enter the object tracking stage. According to the template and the real-time image in the video stream, the Siamese-RPN tracking network will output the bounding box D_2D$_2$ and the confidence C_2. When $C_2 < C_0$C$_2 <$ C$_0$ or the tracking time is longer than the maximum frames, the object detection network will be started again, otherwise, the object tracking will continue. Finally, the algorithm will end when the video stream is loading done. From another perspective, this method can also be regarded as an adaptive template update for object tracking [31].

In the real-time experiment of USV, it is found that the miss rate of the proposed object detection-tracking fusion scheme is greatly reduced compared with the original algorithm. As shown in Fig. 13, the original network may have the wrong detection results. For example, there are double boxes at T = 1.72 moment (which means two bounding boxes

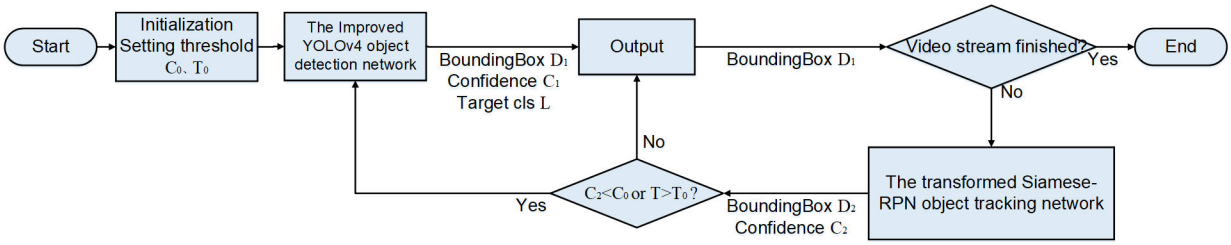

FIGURE 12. The object detection-tracking fusion scheme.

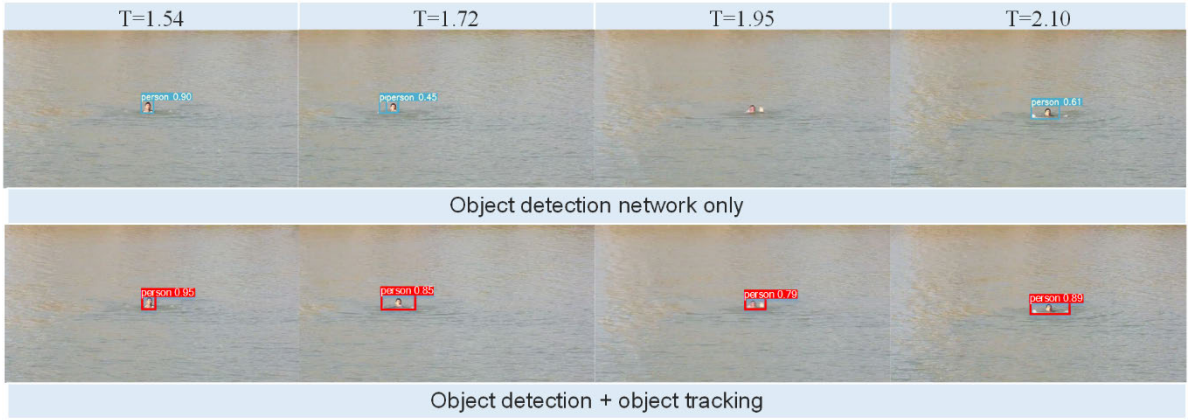

FIGURE 13. Comparison of detection results after introducing tracking network.

detected by the same one person); at T = 1.95 moment the object missed, and at T = 2.10 moment the bounding box has low confidence. However, after integrating the object and tracking method, these bad detections both have been improved. In a word, frame-by-frame object detection cause in miss detection long-time video detection, while the object tracking network can greatly make up for the shortcomings in real-time navigation of USV. These accurately detected and tracked ship bounding boxes will constitute the ship trajectory, which can be used for ship behavior recognition [15] and for promoting the intelligent vessel traffic services [32].

B. COLOR DETECTION BASED ON LC SALIENCY ALGORITHM

For the further needs of information perception, the water target recognition system not only outputs the bounding box but also needs to determine the main color of the target. The key to judging the main color is to separate the foreground, which can isolate the influence of the background color. Therefore, this paper designs a color extraction scheme based on the local contrast (LC) saliency algorithm [33]. The principle of the LC saliency algorithm is to calculate the sum of the distance between each pixel and all other pixels in the image, which will be its saliency value.

Fig. 14 shows the process of color detection. Through the bounding box detected by the network in the early stage, the main region of the object is extracted. Then, according to the LC saliency algorithm, the mask of the foreground can be acquired. Based on the mask, only the foreground color information can be counted, so as to eliminate the interference of background color value.

After the foreground is extracted, the next step is to determine the color category of the foreground region. In this paper, color is divided according to the HSV (hue-saturation-value) color model. In HSV color space, hue represents the color information, in a computer image, whose value ranges from 0 to 180; saturation represents the degree of color approaching the spectral color, whose value ranges from 0 to 255; and the value represents the color brightness, whose value ranges from 0 to 255. There are different opinions on the specific criteria of color division. The color criteria adopted in this paper based on experiments are shown in Table 6.

The last step is shown in the right part of Fig. 14. The original image will be divided into three color layers (H-S-V). According to the judgment basis in Table 6, the color category of each pixel in the foreground area can be judged. By the method of histogram statistics, the largest number of color categories can be obtained as the main color of the water target.

The color detection scheme in this paper can be summarized as follows:

(1) Firstly, according to the bounding box detected by the object detection network, get the main region of color extracting.

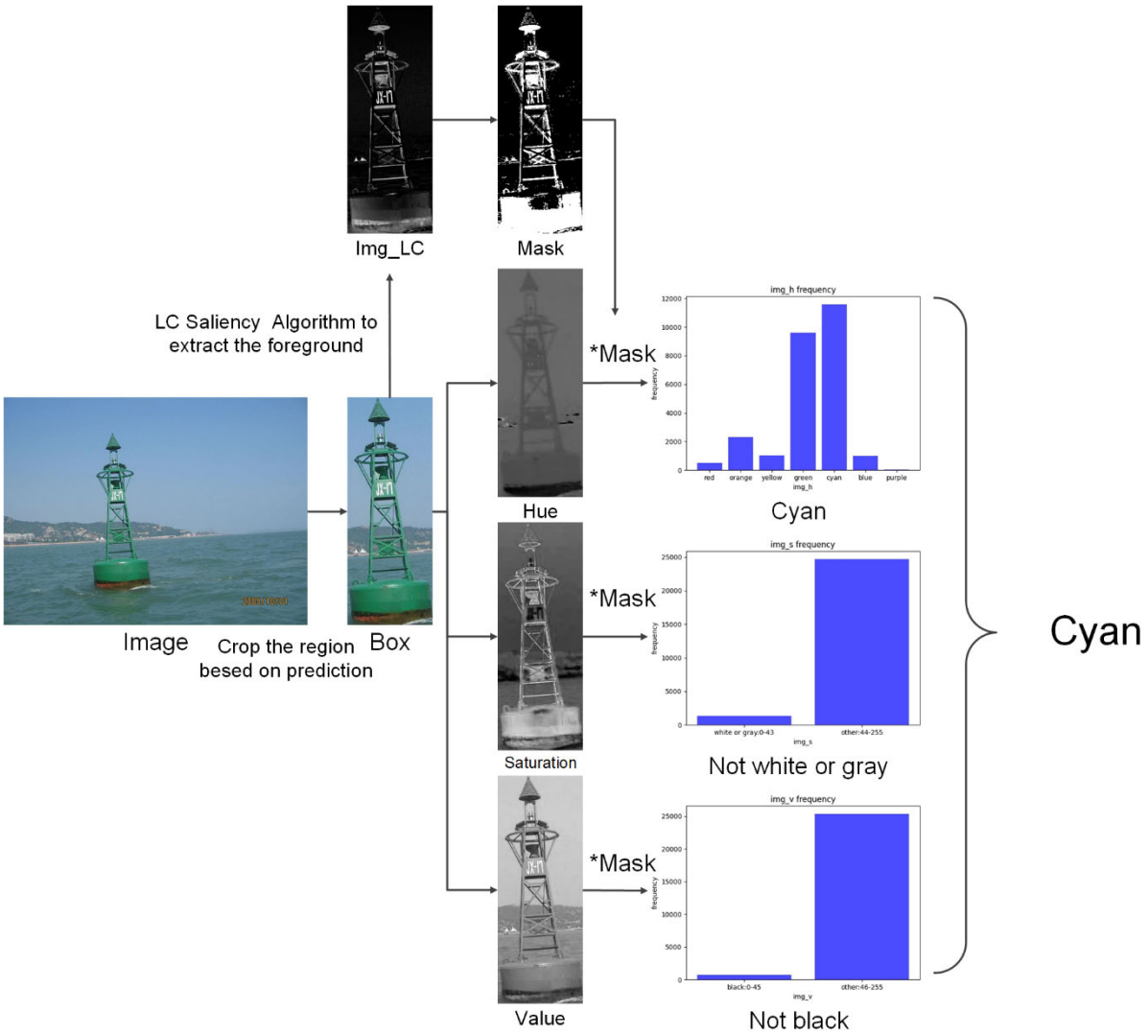

FIGURE 14. The water target color detection algorithm flowchart.

(2) The LC saliency detection algorithm is used to get the foreground mask.

(3) The original image will be separated into H-S-V three layers.

(4) Combining the mask, the foreground area's gray value histogram is calculated.

(5) The space of S and V layers is integrated to judge whether it belongs to black, white, and gray color;

(6) If it doesn't, it will judge which color belongs to which according to the H space, and generally select the most colors in the histogram.

At present, this method has a good detection effect for objects with uniform color distribution, such as navigation buoys, but it is not accurate for objects with complex colors. The frequent histogram statistics and calculation of gray value distance are also time-consuming, so this algorithm can be greatly improved in the future.

VI. THE WATER TARGET RECOGNITION SYSTEM AND EXPERIMENT RESULTS

To build the USV water target recognition system, it is important to deploy the CNN model to an edge computing platform and do some software engineering tasks such as GUI development. Following that, the system was evaluated in a variety of ways, including online learning and offline navigation.

A. THE PROCESS OF DEPLOYMENT

Fig. 15 illustrates the deployment process. It is important to complete model training on a high-performance computer, which is typically accomplished using the efficient PyTorch Deep Learning framework at this stage. It must first be turned into a TensorRT model before being deployed on an edge computing platform. Additional information is available in NVIDIA's official guide. This process may result in a loss of

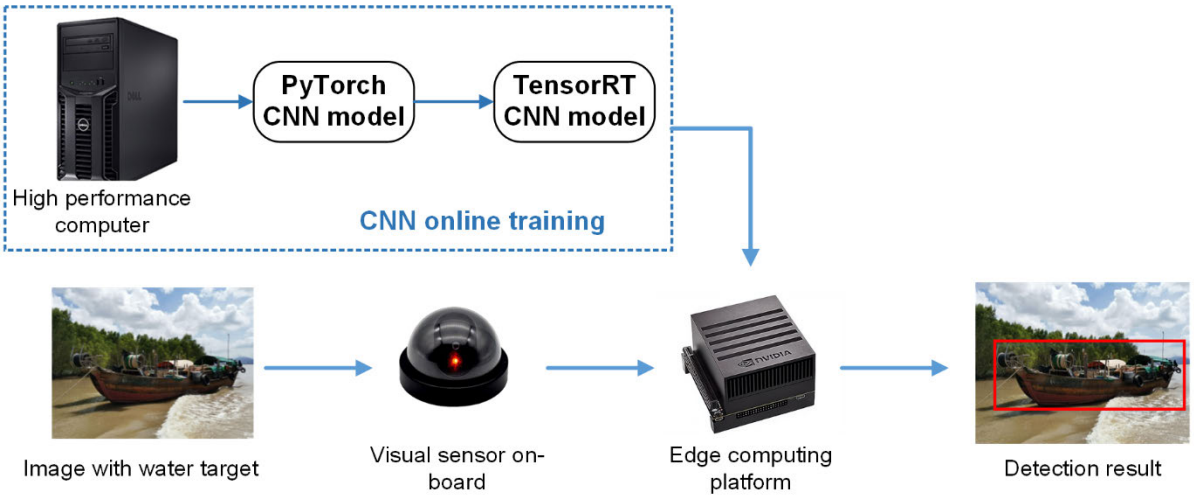

FIGURE 15. The deployment process.

TABLE 6. Color range judgment in HSV space.

Color	Hue	Saturation	Value
Black	0-180	0-255	0-46
Gray	0-180	0-43	46-220
White	0-180	0-30	221-255
Red	0-10/156-180	43-255	46-255
Orange	11-25	43-255	46-255
Yellow	26-34	43-255	46-255
Green	35-77	43-255	46-255
Cyan	78-99	43-255	46-255
Blue	100-124	43-255	46-255
Purple	125-155	43-255	46-255

prediction precision, which is mostly due to the quantization of model weights from floating-point to integer quantities. In general, the NVIDIA Jetson AGX Xavier edge computing platform can be connected to the USV's central control system through a power cable, a connector (typically a USB type-C connection), and an ethernet interface. By invoking the video stream from the visual sensor, the system is able to perform the ship's real-time detection duty for aquatic targets.

B. ONLINE LEARNING RESULTS

Intel i9-7920x CPU, Nvidia RTX2080ti GPU, and Python3.8, OpenCV3.4.2, CUDA10.1, cuDNN7.4, PyTorch1.6, and TensorRT7.0 software environment were used in this experiment.

The water target datasets are separated into training sets of 7452 images and test sets of 2484 images in a 3:1 ratio. The initial learning rate is set to 0.01, the momentum parameter is set to 0.937, the decay parameter is set to 0.000484, and the Focal Loss parameter γ is set to 1.5. The loss function is optimized using the adaptive moment estimating momentum technique (Adam). When the training epoch reaches 286, the network algorithm converges completely.

The final algorithm presented in this article is the outcome of Section IV's pruning, loss function modification, negative sample processing, and picture enhancement. On a high-performance computer, the detection accuracy is 81.74 percent and the inference speed is 69.13 frames per second. After deployment, accuracy reduces to 79.30 percent and computing speed decreases to 26.77 frames per second. Fig. 16 (a) illustrates the AP indicators for each category on the high-performance computer, whereas Fig. 16 (b) illustrates the detection miss rate (b). Overall performance is excellent; even the most difficult barrier targets are accurate to within 69 percent.

C. OFFLINE NAVIGATION

The detection results are presented in Fig. 17 after the system described in this paper is deployed offline. Experiments were conducted on four scenes, including a lakeside pier, an interior river, the Pearl River estuary, and aerial images captured by an unmanned aerial vehicle (UAV). By and large, this method produces a high-quality detection effect that is capable of successfully completing all types of water target detection jobs.

In the scene of a lakeside dock, the man on the shore and the ships swimming in the lake may be spotted correctly and robustly However, it is discovered that the occlusion condition may have a substantial effect, which requires more optimization. We can see that the support pole and antenna wire on the ship significantly blocked the camera during the recognition process for Scene 2 inland river and Scene 3 Pearl River estuary. However, thanks to the blank label training procedure described in Section IV-C, the algorithm accurately and consistently recognizes the navigation buoy

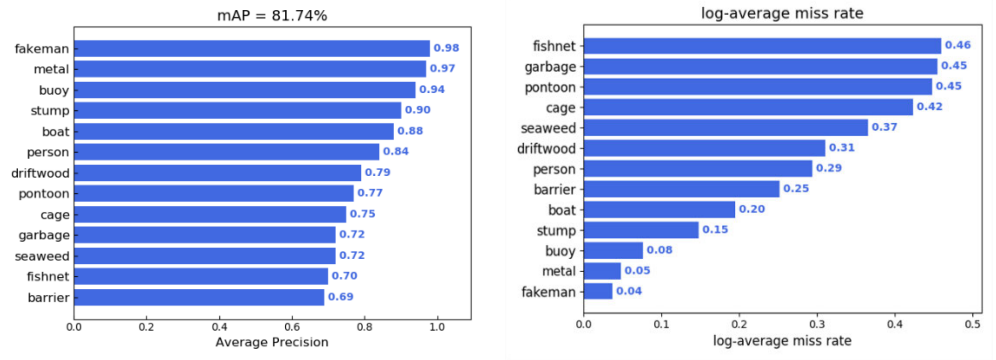

FIGURE 16. The detection results of the improved YOLOv4 algorithm.

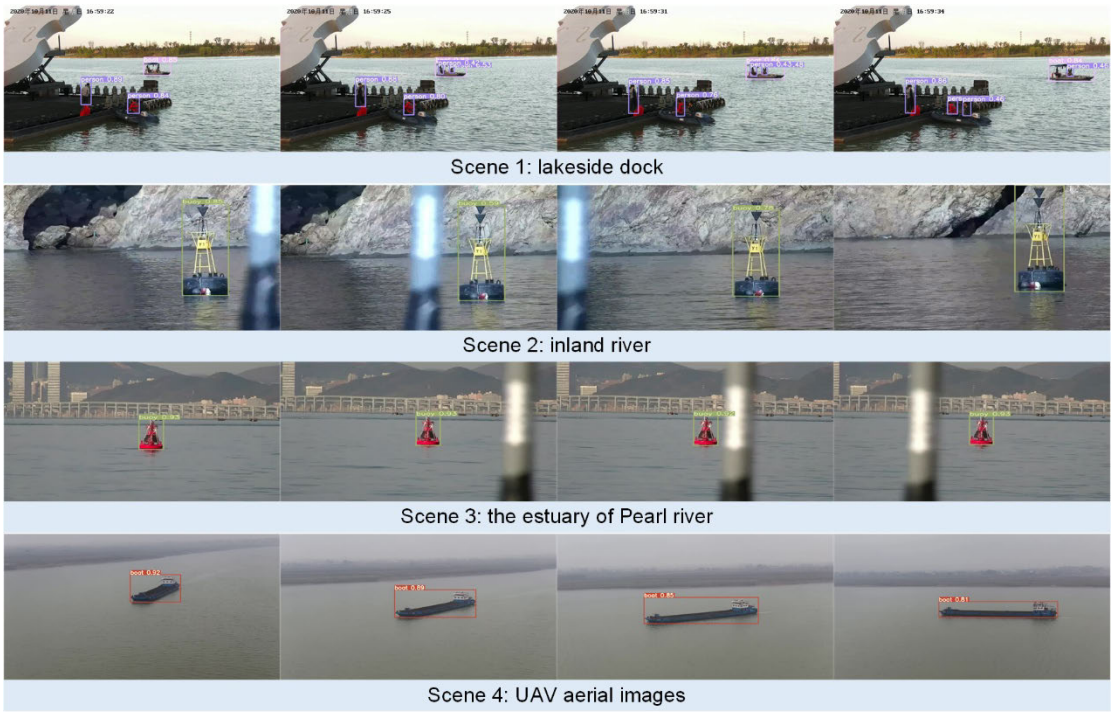

FIGURE 17. The offline navigation of water target recognition system.

in the distance. Additionally, the system detects clearly in Scene 4, whose images are transmitted through UAV from aerial video. Additionally, it is a promising route for future growth in terms of expanding the perception modes of USV.

VII. CONCLUSION AND FUTURE WORK

The purpose of this work is to investigate the lightweight water target detection system based on USVs. To begin, utilizing the human-in-the-loop technique, water target datasets comprising of 9936 images were established. Then, various techniques for optimizing the YOLOv4 algorithm were considered: 1) We pruned the backbone to increase inference speed to 26 FPS from 15 FPS; 2) We used the Focal Loss function to correct for sample imbalance, which resulted in a 0.33% increase in mAP; 3) We used the blank labels training method to reduce false alarms in real-world prediction, which resulted in a 2.02 percent increase in mAP; 4) Several preprocessing methods have been tried to improve detection performance under adverse weather conditions, with histogram normalization. Along with the enhancements to the object detection network, we included a color detection method based on the LC Saliency algorithm for improved obstacle avoidance and cascaded a Siamese-RPN object tracking network to boost the detection stability and real-time performance. Finally, it is deployed on the NVIDIA Jetson AGX Xavier platform for edge computing. The experimental results demonstrate that the water target recognition system is capable of operating in a complicated water target

environment while maintaining its stability, speed, and accuracy.

Water target identification is a rapidly growing field with enormous potential. We describe three intriguing future issues based on what has been accomplished: 1) Several optimizations have been made to YOLOv4's performance, particularly in terms of inference speed, and we look forward to the implementation of a lightweight vision Transformer. 2) Human-in-the-loop annotation improves dataset quality control and speeds up labeling. A more precise and efficient human-machine collaboration strategy will be developed in order to achieve a better balance between annotation speed and quality. 3) In the case of a detection and tracking fusion network, the Kalman filter can be utilized to correlate targets in the front and back frames.

REFERENCES

[1] R. N. Strickland and H. I. Hahn, "Wavelet transform methods for object detection and recovery," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 724–735, May 1997, doi: 10.1109/83.568929.
[2] P. J. Withagen, K. Schutte, A. M. Vossepoel, and M. G. Breuers, "Automatic classification of ships from infrared (FLIR) images," *Proc. SPIE*, vol. 3720, pp. 180–187, Jul. 1999, doi: 10.1117/12.357157.
[3] A. A. Smith, "Identification and tracking of maritime objects in near-infrared image sequences for collision avoidance," in *Proc. 7th Int. Conf. Image Process. Appl.*, 1999, pp. 250–254.
[4] W.-J. Zeng, L. Wan, T.-D. Zhang, and S.-L. Huang, "Simultaneous localization and mapping of autonomous underwater vehicle using looking forward sonar," *J. Shanghai Jiaotong Univ. (Science)*, vol. 17, no. 1, pp. 91–97, Feb. 2012.
[5] J. Shi, J. Jin, and J. Zhang, "Object detection based on saliency and sea-sky line for USV vision," in *Proc. IEEE 4th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Dec. 2018, pp. 1581–1586, [Online]. Available: https://ieeexplore.ieee.org/ielx7/8735656/8740347/08740763.pdf?tp=&arnumber=8740763&isnumber=8740347&ref, doi: 10.1109/itoec.2018.8740763.
[6] Y. Peng, Y. Yan, B. Feng, and X. Gao, "Recognition and classification of water surface targets based on deep learning," *J. Phys., Conf. Ser.*, vol. 1820, no. 1, Mar. 2021, Art. no. 012166.
[7] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018, doi: 10.1109/Tgrs.2018.2848901.
[8] Z. Wei, "The detection and tracking of moving targets with unsteady image for unmanned surface vehicle," M.S. thesis, Dept. Des. Construct. Nav. Archit. Ocean Struct., Harbin Eng. Univ., Harbin, China, 2019.
[9] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/tip.2003.819861.
[14] R. Munro, *Human-in-the-Loop Machine Learning*. Shelter Island, NY, USA: Manning, 2020, p. 456.
[15] X. Chen, L. Qi, Y. Yang, Q. Luo, and O. Postolache, "Video-based detection infrastructure enhancement for automated ship recognition and behavior analysis," *J. Adv. Transp.*, vol. 2020, Jan. 2020, Art. no. 7194342, doi: 10.1155/2020/7194342.
[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37. Jul. 2015, pp. 448–456.
[17] C. Y. Wang, H. Y. M. Liao, I. H. Yeh, Y. H. Wu, P. Y. Chen, and J. W. Hsieh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 390–391.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8691, 2014, pp. 346–361, doi: 10.1007/978-3-319-10578-9_23.
[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007, doi: 10.1109/Iccv.2017.324.
[21] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.
[22] N. Aramideh, A. Hashemi, and W. M. Seiler, "Computing the resolution regularity of bi-homogeneous ideals," *J. Symbolic Comput.*, vol. 103, pp. 141–156, Mar. 2021.
[23] C. Jing and J. Hou, "SVM and PCA based fault classification approaches for complicated industrial process," *Neurocomputing*, vol. 167, pp. 636–642, Nov. 2015, doi: 10.1016/j.neucom.2015.03.082.
[24] H. Lu, Y. Li, T. Uemura, H. Kim, and S. Serikawa, "Low illumination underwater light field images reconstruction using deep convolutional neural networks," *Future Gener. Comput. Syst.*, vol. 82, pp. 142–148, May 2018, doi: 10.1016/j.future.2018.01.001.
[25] M. A. Yousuf and M. R. H. Rakib, "An effective image contrast enhancement method using global histogram equalization," *J. Sci. Res.*, vol. 3, no. 1, p. 43, Dec. 2010.
[26] Z. Li, "Contrast limited adaptive histogram equalization," *Comput. Knowl. Technol.*, vol. 6, pp. 2238–2241, Dec. 2010.
[27] H. Zhang and N. Wang, "On the stability of video detection and tracking," 2016, *arXiv:1611.06467*.
[28] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Comput. Vis. (ECCV) Workshops*, vol. 9914, 2016, pp. 850–865, doi: 10.1007/978-3-319-48881-3_56.
[29] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.
[30] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[31] B. Yan, L. Xiao, H. Zhang, D. Xu, L. Ruan, Z. Wang, and Y. Zhang, "An adaptive template matching-based single object tracking algorithm with parallel acceleration," *J. Vis. Commun. Image Represent.*, vol. 64, Oct. 2019, Art. no. 102603, doi: 10.1016/j.jvcir.2019.102603.
[32] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, "Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6G-enabled maritime IoT systems," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5374–5385, Apr. 2021, doi: 10.1109/JIOT.2020.3028743.
[33] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, Santa Barbara, CA, USA, Oct. 2006, pp. 815–824.