Journal of *Marine Science and Engineering*

MDPI

Article

Improved Convolutional Neural Network YOLOv5 for Underwater Target Detection Based on Autonomous Underwater Helicopter

Ruoyu Chen and Ying Chen *

Ocean College, Zhejiang University, Zhoushan 316021, China; 22134078@zju.edu.cn
* Correspondence: ychen@zju.edu.cn

Abstract: To detect a desired underwater target quickly and precisely, a real-time sonar-based target detection system mounted on an autonomous underwater helicopter (AUH) using an improved convolutional neural network (CNN) is proposed in this paper. YOLOv5 is introduced as the basic CNN network because of its strength, lightweight and fast speed. Due to the turbidity and weak illumination of an undesirable underwater environment, some attention mechanisms are added, and the structure of YOLOv5 is optimized to improve the performance of the detector for sonar images with a 1–3% increment of mAP which can be up to 80.2% with an average speed of 0.025 s (40 FPS) in the embedded device. It has been verified both in the school tank and outdoor open water that the whole detection system mounted on AUH performs well and meets the requirements of real time and light weight using limited hardware.

Keywords: underwater target detection; sonar images; CNN; improved YOLO

Citation: Chen, R.; Chen, Y. Improved Convolutional Neural Network YOLOv5 for Underwater Target Detection Based on Autonomous Underwater Helicopter. *J. Mar. Sci. Eng.* **2023**, *11*, 989. https://doi.org/10.3390/jmse11050989

Academic Editors: Mai The Vu and Hyeung-Sik Choi

Received: 15 April 2023
Revised: 28 April 2023
Accepted: 4 May 2023
Published: 6 May 2023

1. Introduction

Complex underwater environments have overwhelming characteristics, including turbidity and weak illumination, that bring enormous challenges for general object detectors based on common optical RGB images. A considerable amount of work toward target detection algorithms has been put forward with the development of deep learning and convolution neural networks(CNN), starting from LeNet5 [1] in 1990 based on the invention of Neocognitron with Fukushima [2] and boosting from Alexnet [3]. Currently, nearly all advanced detection models, especially for underwater target detection, use the CNN framework, which has already replaced traditional methods such as the digital signal processing based on sound waves [4], statistical features of images, and machine learning such as SVM [5], decision tree [6] and BP neural network [7]. The reason for wide use is that CNNs can produce, extract and fuse hierarchical features automatically instead of involving humans to capture information in varying scales across multiple layers, which allows CNNs to generate robust and distinctive features for accurate detection. In general, there are two types of CNN frameworks. One-stage networks represented by SSD [8] and YOLO series [9–12] are end-to-end detectors with no requirements of a region proposal generation step, while two-stage networks need to form a proposal network to first search for targets, and then, they use a second network to fine-tune these proposals and finally output the detection results represented by RCNN [13], Fast RCNN [14] and Faster RCNN [15], with higher accuracy but slower inference speed. Based on the real-time and light-weighted need of the project, YOLOv5 is chosen as the main framework, and the focus research forms a whole, efficient and high-accuracy detection system mounted on AUH by improving it.

However, traditional CNN detection frameworks are mostly used for optical imagery, and the performance of the underwater optical imaging systems is largely limited by the special environmental conditions of being underwater. In deep or muddy water, we could not obtain one clear image for training and detecting. Thus, we chose a detection

method using acoustic means through sonar, due to it being independent of turbidity and weak illumination and because it has the best ability of using sound waves to transmit information in water. Sonar is more effective than cameras and radar when conducting underwater detection tasks because it can provide more accurate and detailed information. It has already been applied to target detection [16–18], detection based on sonar images has the advantages of higher frequency, longer range, and stronger real time [19,20].

There has been a substantial amount of research within the past few years on underwater target detection based on the improvement of YOLO that is worthy study.

Jing [21] proposed a YOLOv5s-ViT-BiFPN-based neural network to detect damaged houses by introducing the vision transformer into the feature extraction network and by applying BiFPN for multi-scale feature fusion, in which the model increased by 1.23% mAP compared to YOLOv5s. Teerapong [22] developed the transformer-based YOLOX with FPN, which uses ViT as the backbone and adds FPN into YOLOX. Finally, the model surpassed YOLOv5L by 2.56% mAP. Xu [23] proposed the Attention-YOLO using an item-wise attention mechanism, which embedded a channel and spatial attention mechanism in the feature extraction network, in which the model improved at most by 2.5% mAP on the COCO dataset compared with YOLOv3. Experiments by Zhang [24] have showed that the FPN structure of YOLOv4 after using an attention module increased the mAP by 1.08% on the PASCAL VOC dataset. After using depth-wise separable convolution, a multi-scale channel attention module and a modified attentional feature fusion module in MobileNet-YOLOv4, the model obtained a 0.94% increment in mAP and an average of 40FPS speed. Kong [25] created dual-path network (DPN) module and a fusion transition module with YOLOv3 and achieved an improvement of 1.9% mAP on the three-dimensional imaging sonar data. Topple [26] designed a small detector MiNet inspired by YOLO, and MiNet was successfully deployed onboard small autonomous underwater vehicles during a sea trial to detect mines with side-scan sonar images.

Unlike typical RGB images, object detection based on sonar images may cause numerous problems. The first dilemma is the difficulties of capturing sonar images. There is currently not a standard large sonar dataset such as ImageNet [27] and COCO [28] for optical images. Therefore, AUH [29], designed by Zhejiang University, is useful. It is a kind of subsea AUV and is suitable as a dataset because of the characteristics of long-term stable hovering near the seabed and small-scale agile maneuverability. However, acquisition and preprocessing of sonar images have still taken us a great deal of time and cost. Moreover, the sonar images have their own characteristics that make it difficult to extract information. These characteristics include not only low resolution owing to the various kinds of noises and complex underwater sound field but also the ambiguity problem caused by the principle of forming acoustic images that are expressed in two dimensions by projecting three-dimensional images horizontally. In addition, the sonar used in the proposed detection system is cost-effective, has a smaller detection scope and shorter distance, and the images formed have lower resolutions compared to some expansive sonars. Due to the specialties of the sonar images, the design of traditional CNN detection networks may not fit because they are based on and designed for optical images. Thus, the network has to be improved and optimized to help it overcome the poor detection results caused by the shortcomings of sonar images. Adding an attention mechanism is the first thing to improve, and changing the neck and head for structural optimization to better fuse the extracted features and to detect the targets is also verified in the work.

In conclusion, the purpose of this paper is to design a real-time detection system mounted on AUH and to focus on further exploring, improving, and validating algorithms and networks for use in detection underwater, especially seabed targets based on forward-looking sonar images. Improving the detection accuracy and considering speed by modifying YOLOv5 with some attention mechanisms and other structural optimizations, such as BiFPN and decoupled heads, are also explored so that an efficient and effective real-time target detector that is suitable to be mounted on AUH can be built.

In summary, the main contributions of this paper are:

- Some improvements based on YOLOv5 are introduced, including attention mechanisms that add to the backbone, BiFPN that replaces the PANet as the neck, and decoupled heads to separately classify and localize the targets.
- A system for underwater target detection based on sonar images is presented. From training network improved-YOLOv5 to deploying it, to AUH, and then to obtaining the detection results from the algorithm, the whole system to detect the desired underwater targets has been designed.
- Several tank experiments and outdoor tests are implemented to validate the detection system, and the superiority of the improved-YOLOv5 in target detecting is validated.

The remainder of this paper is organized as follows: The real-time underwater target detection system based on sonar images mounted on AUH and its full process with an emphasis on the construction of the improved-YOLOv5 detection model are introduced in Section 2. Especially, the three key techniques, including attention mechanisms, BiFPN, and decoupled heads architecture, are mainly introduced. The applied value of the proposed method is verified by an abundance of experiments in different water environments in Section 3. A comparison between improved YOLOv5 and the original was conducted to prove the advancement of the method. The superiorities and limitations of the proposed method are discussed in Section 4. Finally, the conclusions and the direction of research in the future are drawn out in Section 5.

2. Methods

In order to facilitate the proposed design of the underwater target detection architecture for sonar imagery, the overall workflow of the detection system is firstly described in this section. A brief introduction of YOLOv5 is followed, and then,adding the attention mechanisms, BiFPN and decoupled heads used to optimize the detector.

2.1. Target Detection System Design

The AUH-based underwater target detection system proposed in this paper is based on a typical CNN network, YOLOv5. The network is pretrained with a large number of images from the classical dataset, ImageNet and COCO. The weights and some parameters of this trained network will transfer to the proposed network, which is by improved-YOLOv5 to input some prior information to accelerate the convergence and to make the training process much easier. This process is called transfer learning [30].

Then, the sonar dataset including some common underwater targets, such as shipwreck, anchor, and bucket, collected by the forward-looking sonar in the designed experiments, is used to train the detection model. In order to diminish the network's poor performance and low precision caused by the lack of sonar images, data augmentation tricks such as random rotating, flipping and Mosaic [12] will selectively be used before training. Then, the well-trained model will be deployed through an embedded development board that is loaded onto the AUV to fulfill the mission of real-time detection.

When the detection module obtains the command to begin the detection task, it will continuously send the newest sonar image to the detection network, asking for the result of the algorithm with the frequency of one image per second, and if the detector finds that there is a target, it will return the concrete information of the target to the detection module. The module will combine the information from the detection network and the main controller of AUH to calculate the longitude and latitude of the target, which is called having detected the target. The whole workflow is shown in Figure 1.

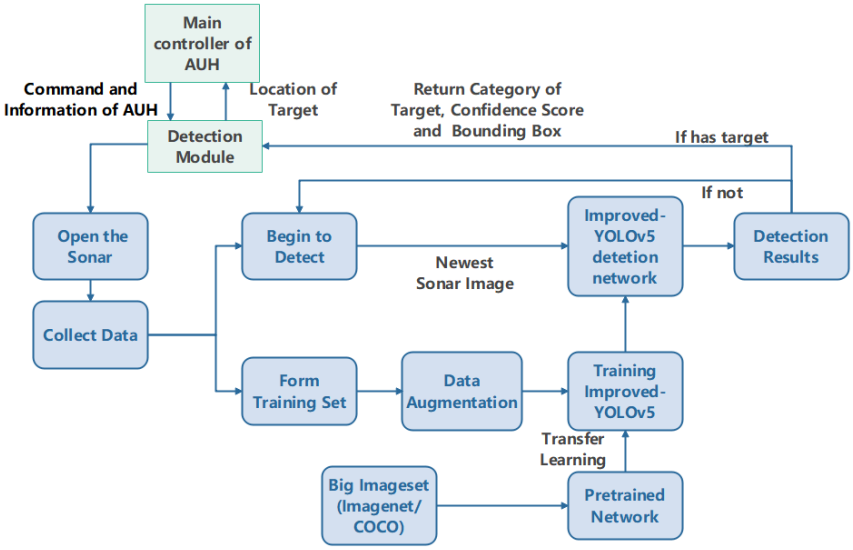


Figure 1. The workflow of the whole AUH-based underwater target detection system.

2.2. Network Building

YOLO is the representation of the one-stage network with no requirements of a region proposal generation step. It takes an input image and divides it into a grid of cells that are responsible for generating multiple bounding boxes and probabilities of each class, and each box comprises the information of object class including X-coordinate, Y-coordinate, width w, and height h, and the confidence score for each object presented in the cell that represents the confidence level of an object in the corresponding bounding box.

2.2.1. Basic Architecture

Basic YOLOv5 will be introduced in brief as shown in Figure 2, and the improved parts compared with YOLOv4 will be the main focus.

Backbone

YOLOv5 uses improved CSPDarkNet53 [31], which consists of C3 blocks as the backbone to extract the image features. C3 with only three convolutions, which is shown in Figure 3, is the evolution version of bottleneckCSP in YOLOv4, and it has fewer parameters and faster speed both in forward calculating and backward broadcasting than bottleneckCSP with the same output. Furthermore, YOLOv5 uses Sigmoid Weighted Liner Unit (SiLU) [32] as the activation function instead of Leaky ReLU [33] because it allows the output range of the network to be between 0 and 1, which makes SiLU perform better than Leaky ReLU in detection applications.

Furthermore, instead of SPP, SPPF is introduced to fuse the global and local features and to obtain the desired size of output without resizing, and SPPF is twice as fast as SPP with the same outcome.

Neck

The neck of YOLO is used for effectively combining the features at different scales. At first, the feature pyramid network (FPN) [34] proposes a top–down pathway to fuse multi-scale features. The strong semantic features of the upper level are passed down, and the entire pyramid is enhanced, but it only enhances the semantic information and ignores the positioning information. Then, PANet [35] adds an extra bottom–up path aggregation network on top of FPN to transmit the positioning information in the lower level. Eventually, the combination of FPN and PANet makes the feature maps of different scales both contain semantic information and positioning information, which ensures the accurate prediction of images of different sizes. The structure is shown in Figure 4, and the different sizes of diamond frames represent the different scales of the feature maps.

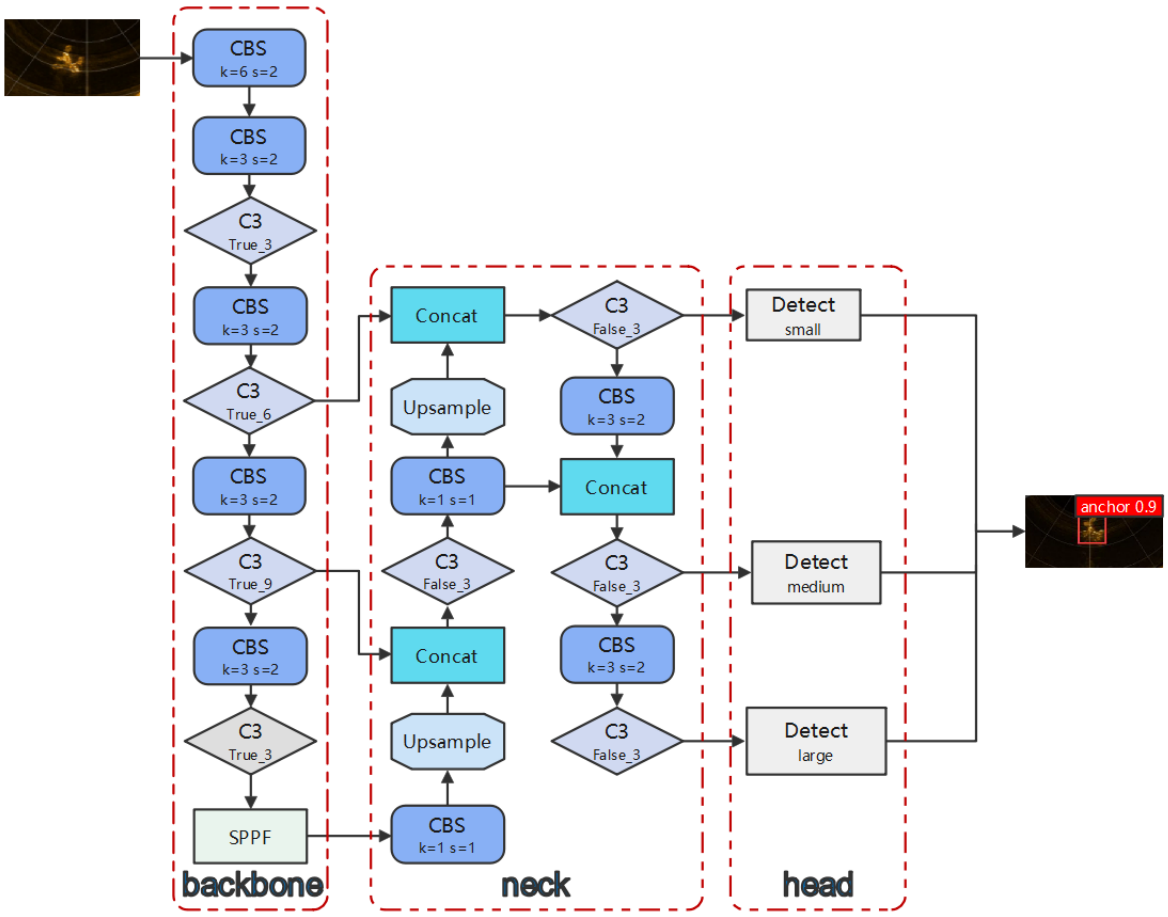


Figure 2. Basic framework: YOLOv5s.

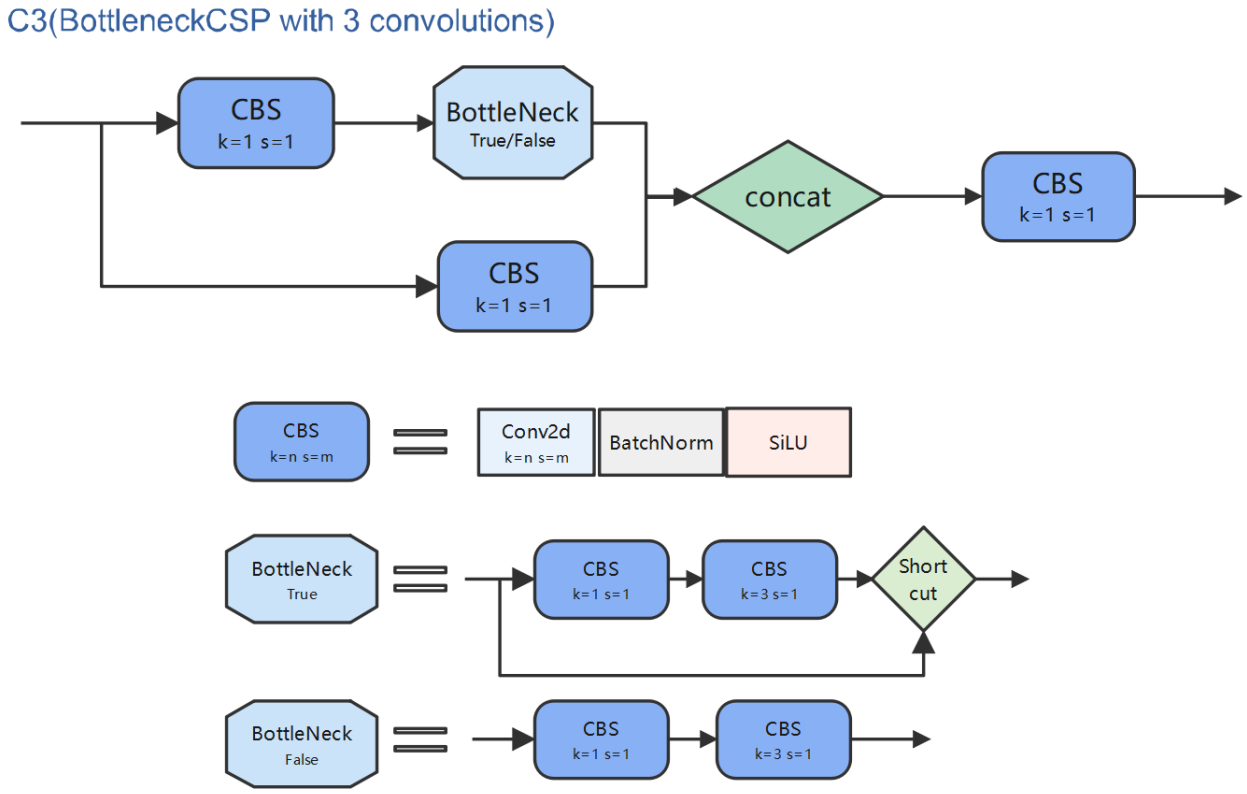


Figure 3. C3 in YOLOv5. The backbone uses C3 (true), and the neck uses C3 (false).

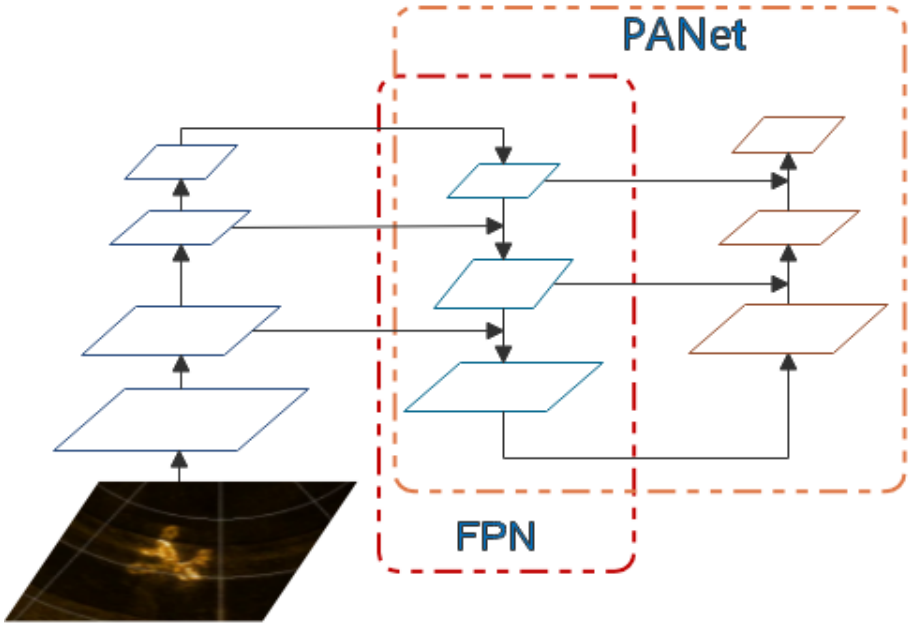


Figure 4. Structure of FPN and PAN.

Detect Heads

Detect heads are employed for final bounding box classification and regression tasks. YOLOv5 uses different scales of heads to separately focus on small, medium, and large targets to make the detection fit for all different sizes of desired targets.

Furthermore, *CIOU* (Complete Intersection over Union)_Loss is used as the loss function of the bounding box instead of *DIOU* (Distance–IoU)_Loss, which is given as:

$$CIOU_{loss} = 1 - IOU + \frac{\rho^2_{(B,B_{gt})}}{c^2} + \alpha\nu \quad (1)$$

where *IOU* (intersection over union) refers to the ratio of the intersection and union area between the predicted and true bounding boxes, $\rho_{(B,B_{gt})}$ represents the distance of the central points of the predicted and true bounding boxes, and c represents the diagonal length of the minimum enclosing rectangle covering the two boxes. α is a weight function, and ν represents the similarity of the aspect ratio.

While *DIOU_Loss* and *CIOU_Loss* are both improvements over traditional *IOU_Loss*, *CIOU_Loss* may be more effective and robust at handling differences in aspect ratio and size variation and can lead to better localization accuracy because of the aspect ratio factor.

Attention Mechanisms

The goal of adding attention modules to the basic YOLOv5 is to enable the network to selectively focus on the most relevant features within an image, while ignoring irrelevant or redundant information.

The structures of CBAM, CA, and GAM are shown in Figure 5, and their ability to enhance the attention to the specific area and target has been verified.

CBAM [36]

CBAM (convolutional block attention module) comprises two blocks, which are channel attention module (CAM) and spatial attention module (SAM). CAM focuses on feature channel weights and adjusts them dynamically to emphasize the most informative channels, and SAM focuses on the spatial relationship of features and adaptively rescales them. CBAM combines these two attention blocks by first applying CAM to the feature map and by then applying SAM to the resulting output.

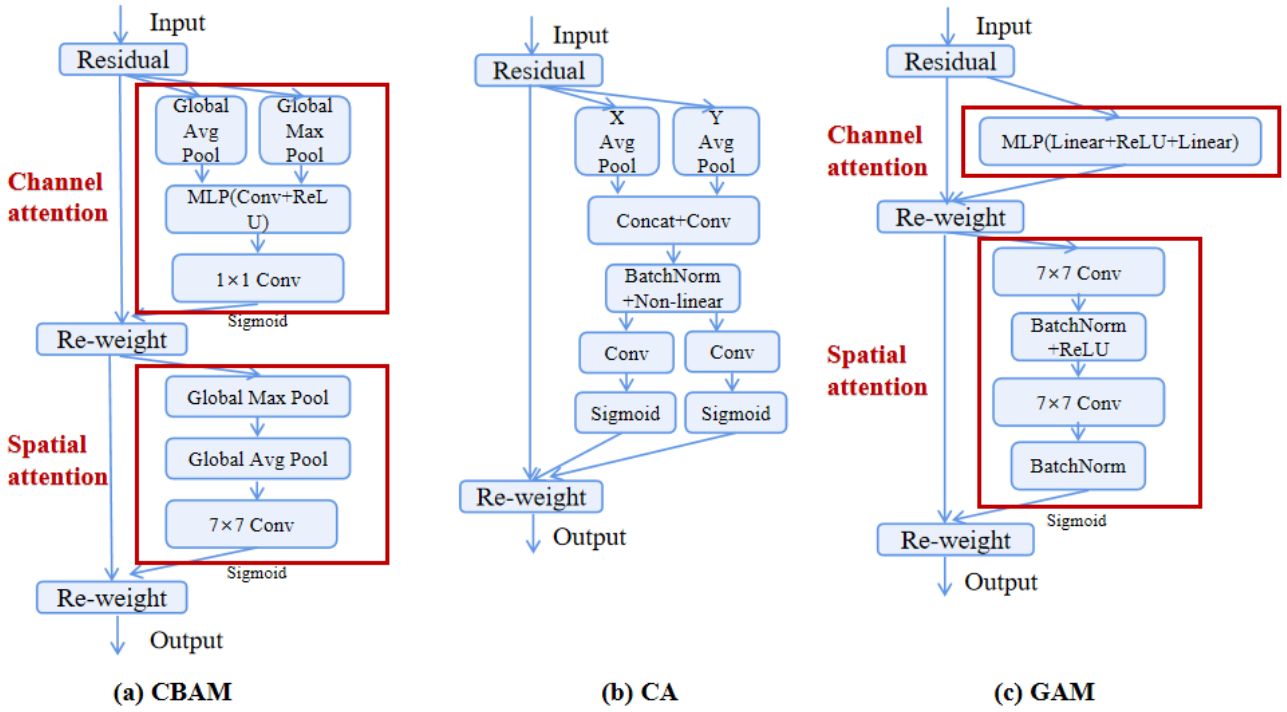


Figure 5. Three attention mechanisms: CBAM, CA, GAM.

By selectively emphasizing both informative feature channels and spatial locations, CBAM can improve the feature representation capabilities of a CNN, leading to better accuracy on the target detection tasks.

CA [37]

Compared to the juxtaposition of spatial and channel feature maps in CBAM, CA (coordinate attention) embeds the position information into the channel attention. CA skillfully combines channel attention with spatial attention so that the remote dependency can be captured along one spatial direction and so that accurate position information can be preserved along the other spatial direction. The resulting feature maps are then encoded separately into a pair of attention maps that are separately sensitive to direction and position and that can be added to the original input feature map to enhance the network's ability to represent targets of interest within the images.

GAM [38]

Although CBAM combines the information of channels and spatial locations, it ignores the interaction of channels with space, thus losing cross-dimensional information. Thus, GAM improves channel attention and spatial attention modules using 3D permutation and MLP to reduce information diffusion and to strengthen global interactions. GAM can emphasize important information and features in all three dimensions.

2.2.2. Other Tricks

BiFPN

BiFPN (bidirectional feature pyramid network) [39] introduces learnable weights to learn the importance of different input features, which are used to build better feature pyramids for object detection by fusing information across multiple resolutions and scales. This is achieved through a series of repeated top–down and bottom–up pathways that work together to generate a set of fused features.

Compared to PANet using a two-stage approach through the use of a top–down and bottom–up pathway, each subsequent BiFPN layer fuses multi-scale feature maps to produce refined representations. BiFPN is a more simple and efficient way to combine multi-

scale features from different backbone levels with fewer parameters and computations and can achieve better results, especially for small targets, as shown in Figure 6.

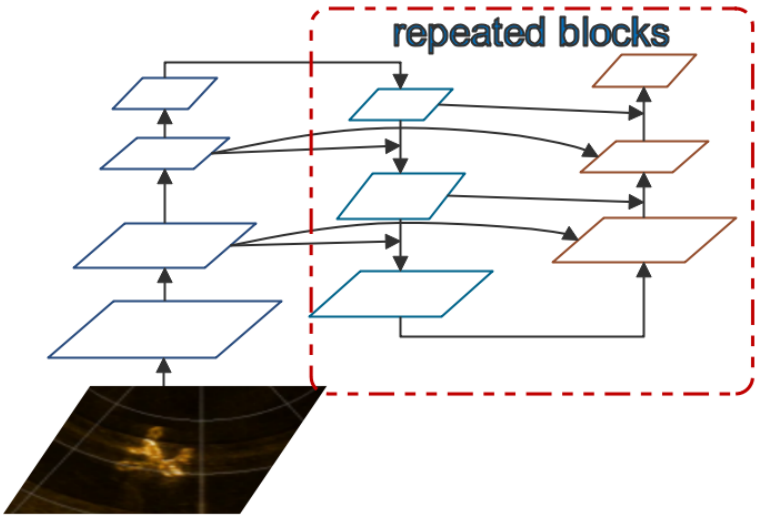


Figure 6. Structure of BiFPN.

Decoupled Heads

In YOLOv5 and other YOLO series, there is a single output layer that computes the position and name of the target together. In contrast, decoupled head architecture [40] separates the classifications into their own respective loss functions and the localization tasks into their own independently trained heads, and then, mixes them together. This separation allows for better optimization and tuning of each task. In general, decoupled heads lead to better performance and faster convergence. Additionally, this architecture is more flexible, allowing for the use of different encoders or backbones and decoders or heads for different tasks.

The differences between coupled heads and decoupled heads that we inserted into YOLOv5 is distinctly noted according to Figure 7.

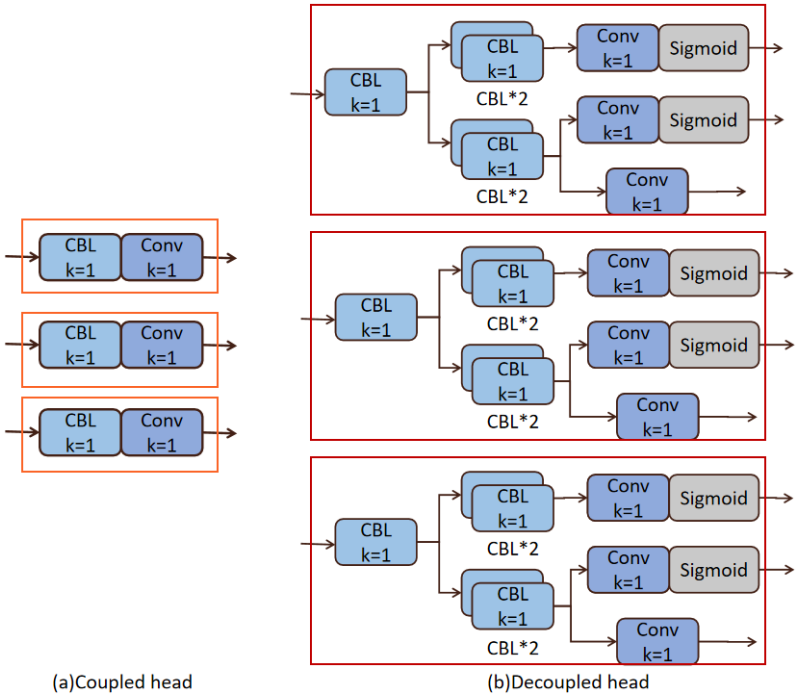


Figure 7. Differences between traditional coupled heads and decoupled heads.

3. Experiments

The experiments are described in detail, from network training and comparison of the models, to the tank experiments and outdoor trials in practice, in this section.

3.1. Detector Training

1. Datasets

Four different datasets were collected in different water environments, including the tank at Ocean College, Zhejiang University, and the lake in Zhoushan, and they were divided into eight common underwater targets: anchor, basket, pillar, shipwreck, human, aircraft, bucket, and uncertain target.

An anchor, a basket, and a pillar-shaped target were placed inside the school's experimental tank. Oculus BluePrint MD750d forward-looking sonar mounted on a ROV was used for collection, named Dataset 1, and it includes 2239 images formed by the software ViewPoint, which the sonar company provided. Dataset 2 is an open-source dataset called the Sonar Common Target Detection Dataset (SCTD) [41], which has 357 images in total, including shipwrecks, humans, and aircraft. SCTD was composed of forward-looking sonar (FLS) images, side-scan sonar (SSS) images and synthetic aperture sonar (SAS) images. This dataset was used for training with the expectation that the model can determine all the types of sonar images to prevent overfitting. Datasets 3 and 4 were collected using MD750d forward-looking sonar, but the experimental site was moved outside. An anchor, a basket, and a group of buckets were place into the lake, and AUH was used to collect data. The program was written by us to obtain the original gray sonar images.

Figure 8 shows some examples of the datasets.

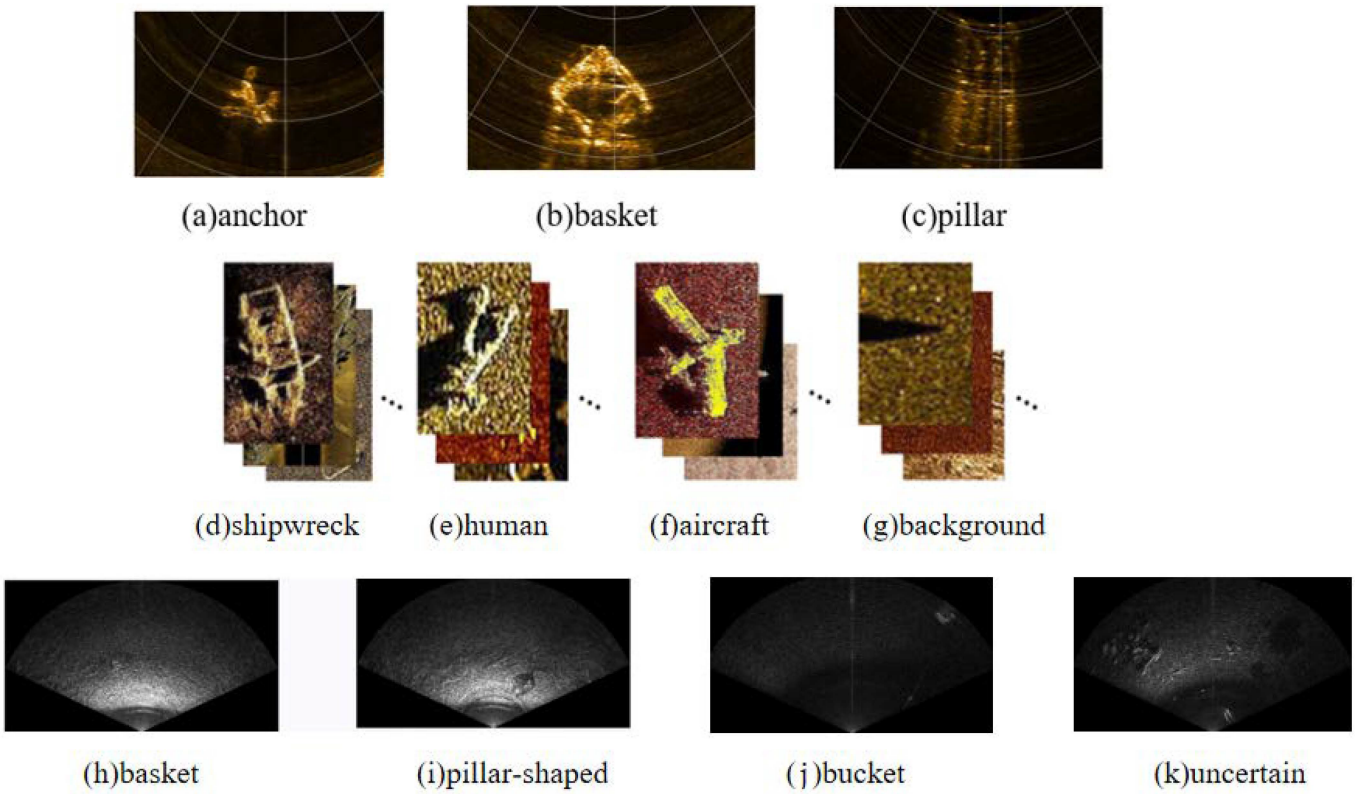


Figure 8. Examples of datasets. The photos in the first row were captured by us, including the anchor, basket, and pillar, which are from Dataset 1. The second row is from SCTD, including shipwrecks, humans and aircraft, which are from Dataset 1. The third row was captured from the lake, which is from Datasets 3 and 4.

2. Hardware Platform
The autonomous underwater helicopter (AUH) [29] is one of the newly developed autonomous submersibles by Zhejiang University, with a disc-shaped design that enables ultra-mobility movements underwater, including full-circle rotation, stationary hovering, and free take-off and landing. It possesses various features, such as small-scale agile maneuverability, long-distance navigation, closed exterior, low operational resistance, and high structural stability. It can cruise for a long time at a fixed height close to the bottom of the water. Named as an underwater helicopter due to its similarity to a land-based helicopter in terms of characteristics, it can be employed as the ideal equipment for target detection or operation at a specific location by performing flexible and agile movements. These specialties make it highly suitable for solving the underwater target detection problem.
Thus, forward-looking sonar was assembled at the front of the AUH at 20 degrees downward, as shown in Figure 9. The data were collected at a height of 5 m away from the bottom of the water.

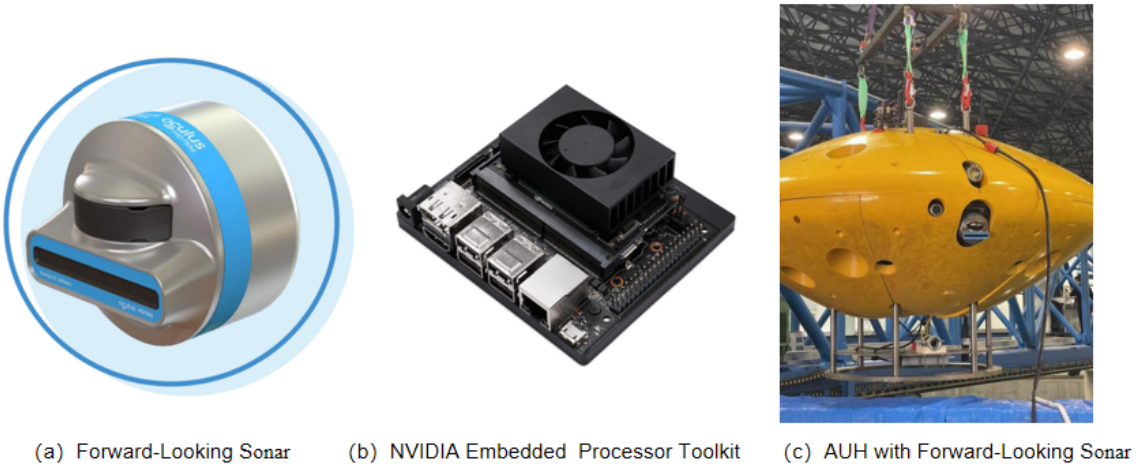


Figure 9. The hardware. (**a**) Forward-looking sonar used in the experiments, which is 4 km pressure-resistant. (**b**) NVIDIA's embedded processor toolkit carrying the detection network model. (**c**) AUH carrying the sonar.

3. Software Platform
The computer configuration and coding environment for training the network are shown in Table 1.

Table 1. Software Platform.

Items	Version
CPU	AMD Ryzen 7 5800 8-Core Processor
GPU	NVIDIA GeForce RTX 3070 Ti
Video memory	8 GB
RAM	16 GB
CUDA	CUDA v11.0 CuDNN v8.0.4
Python	3.7.11
Pytorch	1.7.0
Operating system	Windows11
NVIDIA's Embedded Processor Toolkit	Jetson Xavier NX

4. Parameters
The parameter is the important factor influencing the results of the detection network, and a better outcome will be obtained simply by tuning the parameters. Thus, the standard and unified parameters shown in Table 2 are utilized to compare the networks with equity.

SGD and Adamw have been separately tried as the optimizer to train the models. The outcome shows that Adamw is preferred over SGD due to its ability to converge faster and to handle sparse gradients better. The learning rate is adjusted automatically based on the historical gradients of the parameters and incorporates momentum to help accelerate convergence when using Adamw; on the other hand, SGD only uses a fixed learning rate for all parameters. Warm-up and cosine annealing are the commonly used strategies that also work for adjusting the learning rate during the training to let the model converge faster and better.

Table 2. Training settings.

Training Parameter	Value
train:val:test	8:1:1
batch_size	8 for training and 4 for validating
epochs	300
input_size	(640, 640)
momentum	0.937
weight_decay	0.0005
initial learning rate (lr_0)	0.01 (SGD) 0.001 (Adamw)
cyclical learning rate (lr_f)	0.1 (Cosine annealing)
warmup_epochs	3
warmup_momentum	0.8
warmup_bias_lr	0.1
default anchor size	[32, 31, 47, 46, 65, 60] [59, 75, 83, 75, 84, 87] [97, 98, 117, 114, 118, 144]

5. Training strategies
Some powerful training strategies in YOLOv5 are utilized to encourage the model to learn more context, generalize better to unseen data, and improve model robustness that can handle target variations across multiple scenarios.

- AutoAnchor:
The k-means clustering algorithm is used to self-adaptively generate prior anchors by using all detection frames in the dataset before each training to enable the detection network to obtain more prior knowledge of the underwater target.
- Multi-scale training and distortion:
The input images are randomly resized to different scales during each iteration of the training process. Multi-scale training is commonly used in conjunction with data augmentation techniques such as random cropping, rotating, scaling, flipping, translating, and shearing to geometrically distort the images. The purpose is about exposing the model to objects at different sizes and resolutions and to force it to learn spatial invariance and more robust features that can handle object variations across multiple scales and different water environments.
- Letterbox resize:
Before training the images, all images need to be resized to fit into a fixed size without stretching or distorting the shape so that they can be fed into the neural network. Black bars are added to the top and bottom (or left and right) to fill the empty space created by the new size by using the least amount of bars. This strategy ensures that the relative scale and aspect ratio of objects within the images are preserved and that it can improve the accuracy and accelerate the training compared to the old version of YOLO because of the reduction of the filled area, which is redundant information.

3.2. Detector Evaluation

Before presenting the results of the experiments, the metrics used to evaluate the accuracy of the designed network need to be clarified. The correlated concepts are shown in Table 3:

Table 3. Confusion matrix.

Confusion Matrix		Results from Detection Network	
		Positive	**Negative**
Ground Truth	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The concept in Table 1 is often used to define the detection evaluation indexes. To understand, taking the detecting bucket as an example, TP and TN mean that the results given by the detector are consistent with the ground truth, which is correct. FP means that the network predicts the bucket but that it is actually not there, and FN represents that the detector misses the bucket.

The equations of precision and recall are shown in Equation (5).

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{2}$$

Mean average precision (mAP) is an important metric that is the area under the P-R curve, which is the mean value of AP of all classes under a fixed intersection over union (IOU) threshold, which can be calculated by:

$$\text{mAP} = \frac{\sum_{i=1}^{n} \int_0^1 P(R) dR}{n} \tag{3}$$

mAP_0.5:0.95 means calculating the average mAP from a mAP of 0.5 to 0.95 in intervals of 0.5, where n is the number of total target classes.

F_β also takes into account both precision and recall, which can reflect the relationship between P and R by

$$\frac{1}{F_\beta} = \frac{1}{1 + \beta^2} \times (\frac{1}{P} + \frac{\beta^2}{R}) \tag{4}$$

β can be set to different values to balance the importance of precision and recall. Setting it to 1 stands for the harmonic mean of precision and recall, which is

$$F_1 = \frac{2PR}{P + R} \tag{5}$$

In conclusion, precision, recall, mAP, parameters of the models, and inference time are applied to quantitatively evaluate the object detection performance of the proposed network based on YOLOv5 on the test set.

3.3. Results and Analysis

3.3.1. The Tank Experiments Stage

First, Dataset 1 from the school tank, which is shown in Figure 10, and Dataset 2 from an open source were used to train the model.

Figure 11 shows an example of the progress of the placement of the target and detection using AUH.

The purpose of target detection is to predict a set of bounding boxes and category labels for each desired target. The smallest version of YOLOv5 was chosen because of its speed and applicability to real-time detection. The visualized result of YOLOv5s is shown in Figure 12, and the shortcomings of YOLOv5s can be intuitively noted, which easily miss the target. It missed the ship in the left corner when validating because of the lack of ship samples.

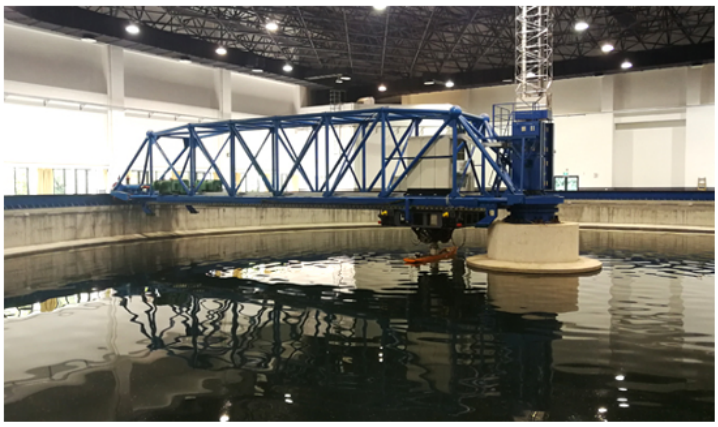
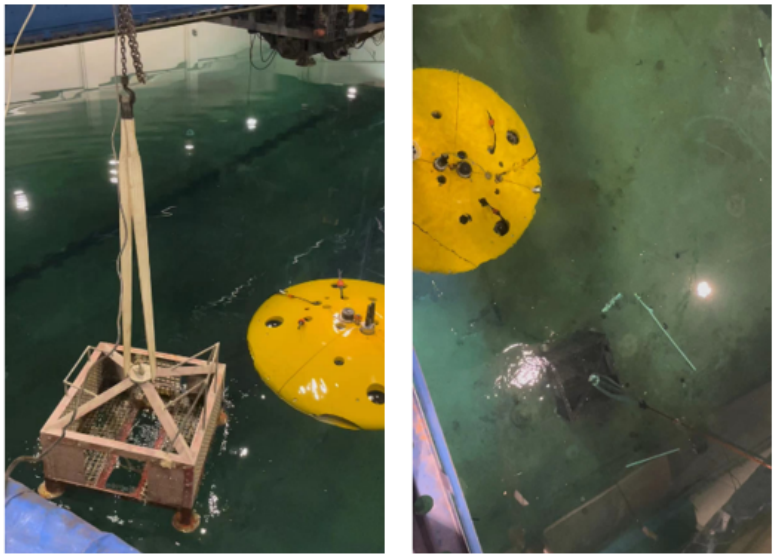
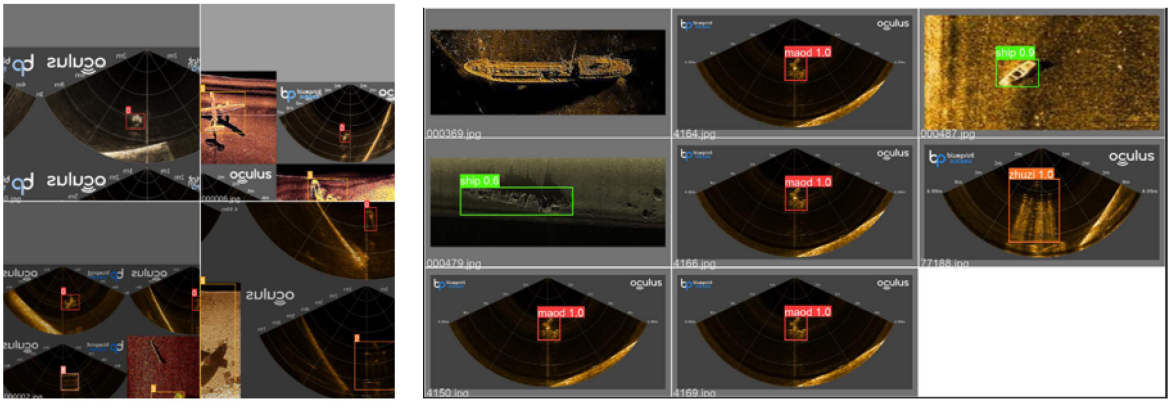


Figure 10. The experimental tank at Ocean College, Zhejiang University.



(a) Placement of theTartet (basket) (b) AUH Approaching the Tartet (basket)

Figure 11. (**a**) A basket is placed into the tank. (**b**) AUH approaches the target to detect.



(a)training (b)validation

Figure 12. (**a**) Training process. The strategies of data augmentation mentioned before, such as flipping and mosaic, are shown here. (**b**) Validation process, which is the outcome when the model meets unseen images.

Thus, we tried to add the attention mechanisms, BiFPN, and decoupled head structure to the original YOLOv5s to improve accuracy. The qualitative results are shown in

Table 4, which shows that CA achieved the best outcome, increasing mAP by 2.1% without many parameters added, and it even had the least parameters and weights among these three mechanisms.

Table 4. Qualitative comparison among baseline, CBAM, CA, and GAM.

Structure (Datasets 1 + 2)	YOLOv5s Baseline	YOLOv5s+ CBAM	YOLOv5s+ CA	YOLOv5s+ GAM
Backbone	C3 C3 C3 C3	C3 C3 C3 C3 + CBAM	C3 C3 + CA C3 + CA C3 + CA	C3 C3 C3 C3 + GAM
Neck	C3 C3 C3 C3	C3 C3 + CBAM C3 + CBAM C3 + CBAM	C3 C3 C3 C3	C3 C3 + GAM C3 + GAM C3 + GAM
mAP(0.50:0.95)	0.769	0.77 (+0.1%)	**0.79 (+2.1%)**	0.788 (+1.9%)
Layers	213	257	237	257
Parameter (M)	7.02	7.10	7.06	11.05
Weights (MB)	13.7	13.9	13.8	21.4

The networks' outcome before and after improvement through the curves can be directly compared in Figure 13.

BiFPN and decoupled heads can both improve the accuracy significantly by 2.1% and 3.3% according to Table 5, which can be seen intuitively in Figure 14 and which can converge faster than the original YOLOv5, as Figure 15 shows. Decoupled heads can boost mAP to 80.2%, which are the greatest detection results, but the structure also doubles the parameter and has longer inference time, which is not beneficial for real-time detection. The detection speed is another factor that needs to be considered cautiously besides accuracy. The balance between these two factors decides the model's feasibility in practical applications in an open water area.

Table 5. Qualitative comparison among baseline, BiFPN, and decoupled heads.

Structure (Dataset 1 + 2)	Baseline	+BiFPN	+Decoupled Heads
MAP (0.5:0.95)	0.769	**0.79 (+2.1%)**	**0.802 (+3.3%)**
Layers	213	229	267
Parameter	7.02 M	8.09 M	14.30 M
Inference Time (ms)	3.5	4.0	5.4
Weights (MB)	13.7	14.8	27.7

Intuitively, regardless if an attention mechanism, BiFPN or decoupled heads are added, the ship can be detected with a relatively high confidence score, as shown in Figure 16.

3.3.2. The Lake Test Stage

The lake test, located in the west of Zhoushan Island, is named Golden Bay, and it has an average water depth of 7 m. The test site and the offshore AUH deployment are shown in Figure 17, which are near the coordinates (121.944° E, 18.174° N). Two lake tests were conducted in June and November 2022 in Golden Bay for and Dataset 3, which includes an anchor and basket, and for Dataset 4, which includes groups of buckets. In total, 2937 sonar images were collected. In addition, an uncertain class was added to annotate some underwater objects that could not be recognized.

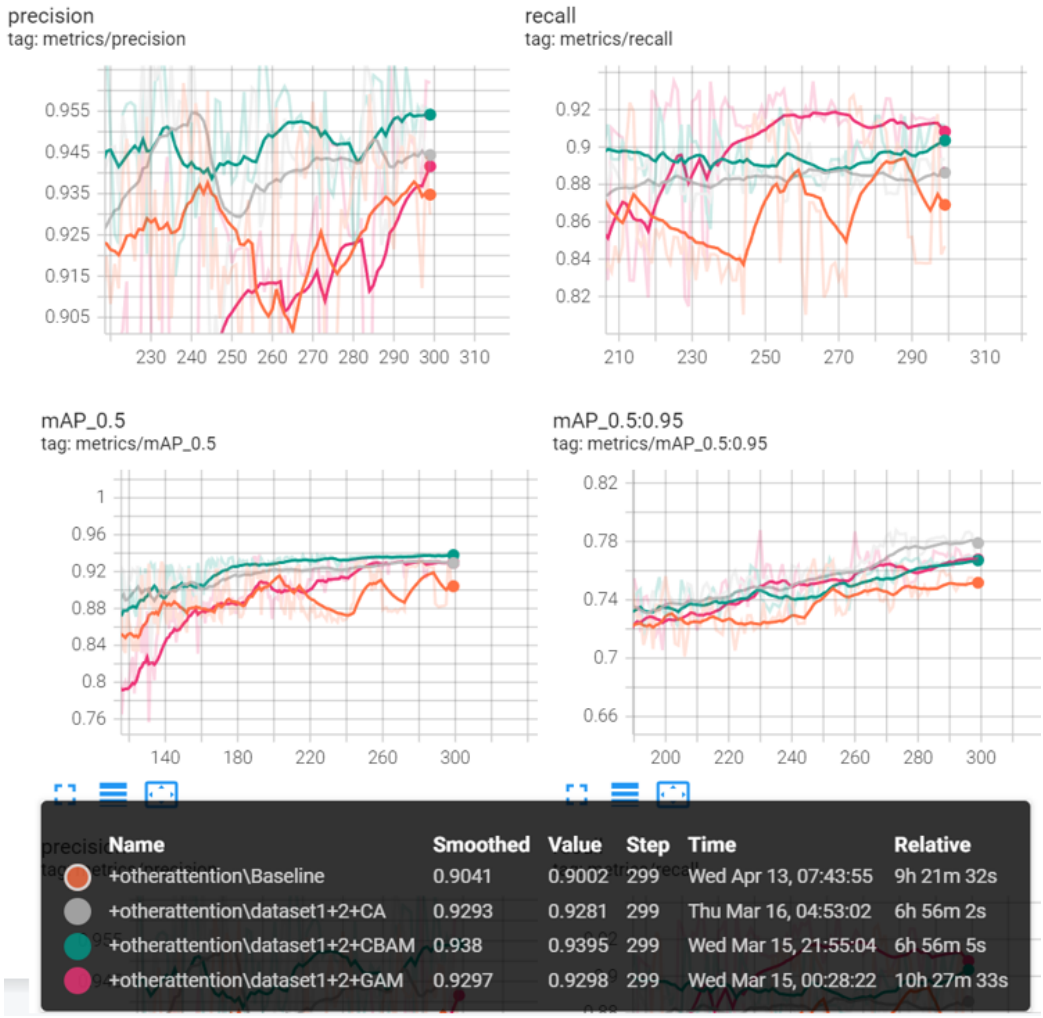


Figure 13. Precision, recall, mAP (0.5) and mAP (0.50:0.95) among baseline, CBAM, CA, and GAM.

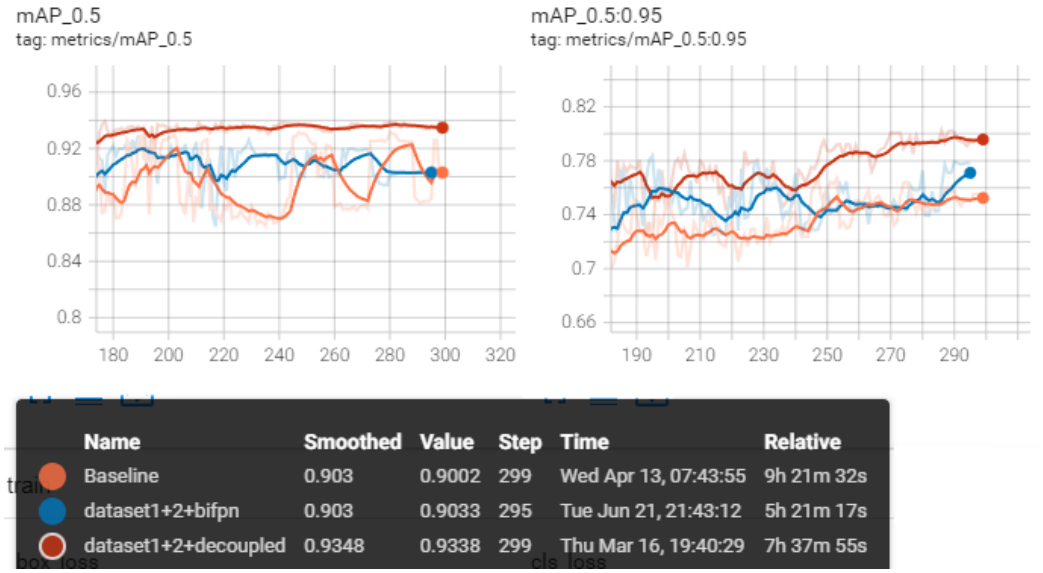


Figure 14. MAP (0.5) and mAP (0.50:0.95) among baseline, BiFPN, and decoupled heads.

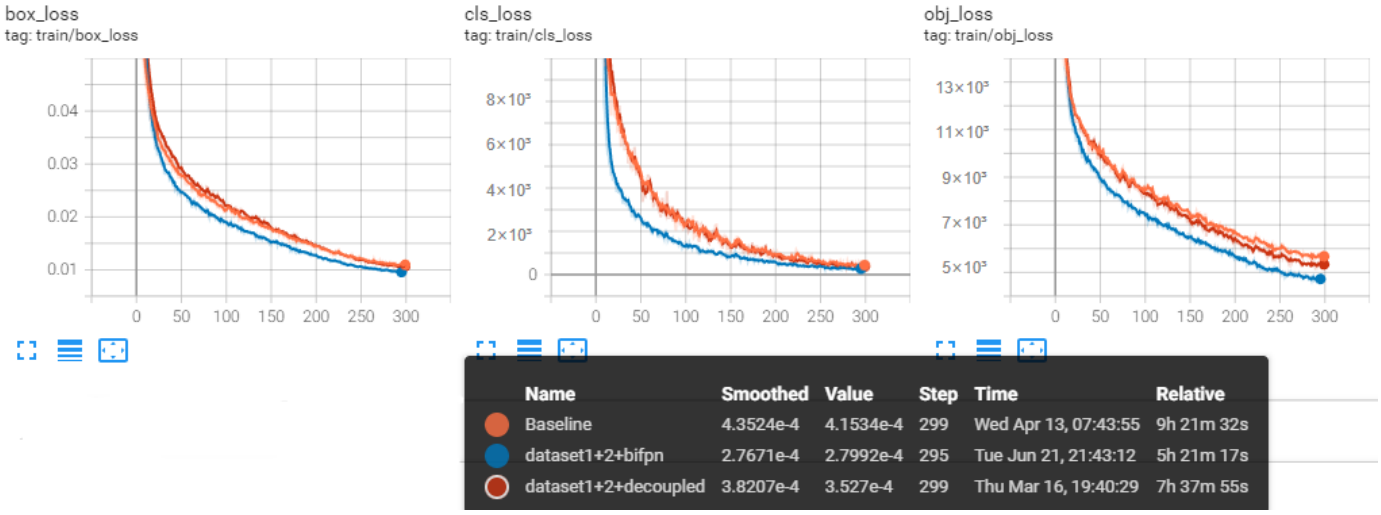


Figure 15. Loss comparison among baseline, BiFPN, and decoupled heads.

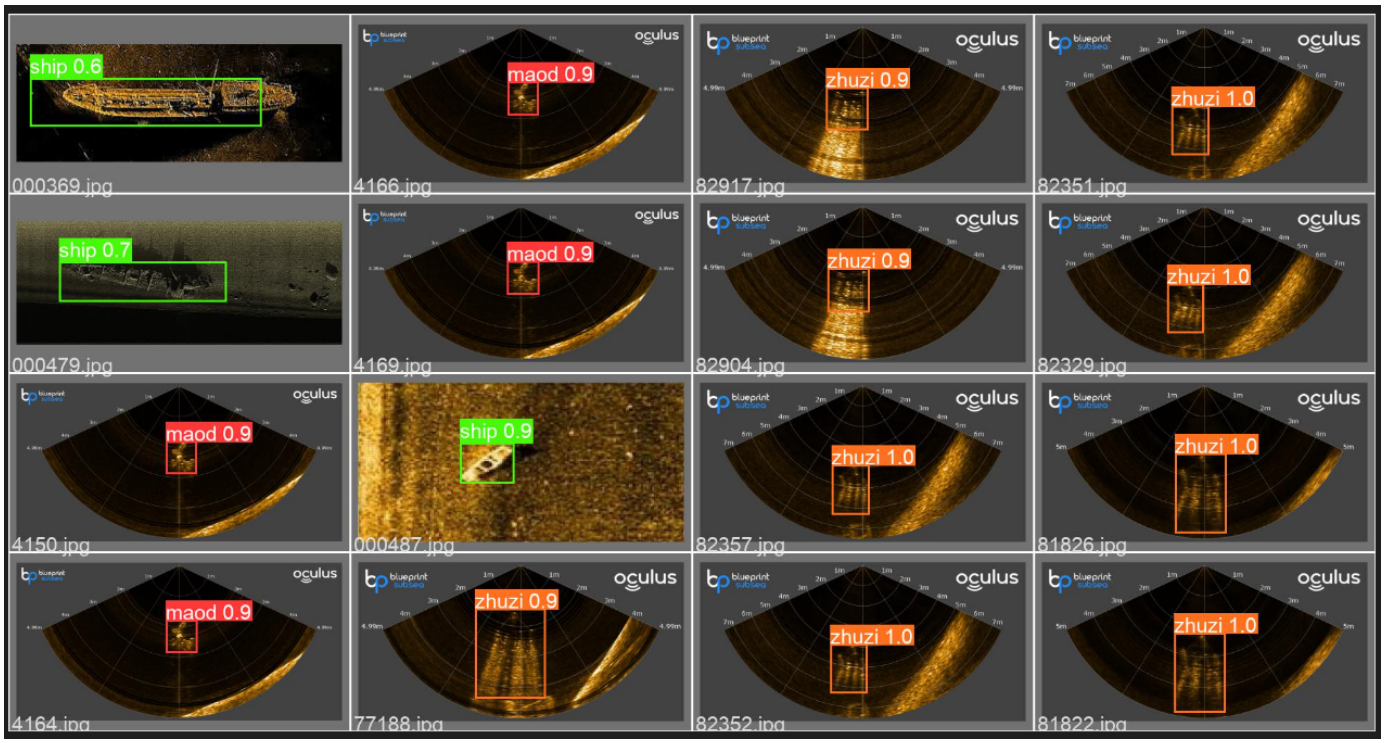


Figure 16. The ship was detected by improved YOLOv5.

Figure 17. The scenery of Golden Bay and AUH offshore deployment.

First, Dataset 3 was added into training to prove the effectiveness of the proposed attention mechanisms with no new targets joining. According to the results shown in Table 6, GAM outperforms a little more than the other two blocks. Different datasets have different optimized attention mechanisms, but they all can improve the mAP to some degree while barely increasing the inference time.

Table 6. Comparison among baseline, CBAM, CA, and GAM.

Structure (Datasets 1 + 2 + 3)	YOLOv5s Baseline	YOLOv5s+ CBAM	YOLOv5s+ CA	YOLOv5s+ GAM
Backbone	C3	C3	C3	C3
	C3	C3	C3 + CA	C3
	C3	C3	C3 + CA	C3
	C3	C3 + CBAM	C3 + CA	C3 + GAM
Neck	C3	C3	C3	C3
	C3	C3 + CBAM	C3	C3 + GAM
	C3	C3 + CBAM	C3	C3 + GAM
	C3	C3 + CBAM	C3	C3 + GAM
mAP (0.50:0.95)	0.763	0.783 (+2.0%)	0.773 (+1.0%)	**0.784 (+2.1%)**
Inference Time (ms)	3.5	3.7	4.3	6.1

Then, the network was trained to recognize the new target, the groups of the buckets. Especially, the best training weight of Datasets 1 + 2 was used as the pretrained weight to converge the new training faster and to improve the accuracy. The network was embedded in the NVIDIA NX toolkit, and the speeds of different formats of weight files using one NVIDIA 3070Ti were compared and are shown in Table 7. Finally, the fastest method, which used the TensorRT framework to turn the weights into engines and then used the Python file to accelerate the inference time, was chosen. The improved YOLOv5s with the attention mechanisms in this acceleration method was finally used. Eventually, the model can achieve an average speed of 0.025 s (40 FPS) in the embedded device and fulfill the real-time requirements.

Table 7. Inference time comparison among different formats of weights.

Methods	.pt	.onnx	.engine (with C++)	.engine (with Python)
Average Inference Time (ms)	55	1311	30	**21**

Especially, this was the first time the proposed detection models, using improved-YOLOv5 on AUH to form the detector and to fulfill the whole detection procedure, were tested. AUH executed an area of 100 × 100 m scanning task centered around the targets' coordinates at a fixed height of 5 m with an average speed of 1.0 kn. The aim of the experiments was to detect groups of buckets placed in the lake.

The results shown in Figure 18 given by the real-time detection network are correct and have proven that the CNN-based detection model proposed in this paper can detect the underwater bucket target well.

In addition, some uncertain objects were detected, as shown in Figure 19, which might have been the rock or mud pile at the bottom of the lake.

The main controller of AUH sent a command to the detection module using TCP to open the sonar and to begin the detection task, and then, the module sent the newest sonar image to the algorithm to ask for the result. If the algorithm recognizes the bucket, it will send back the category of the detected target, confidence score, and the XY coordinate values. The detection module will combine the target's information from the algorithm

and the information of the AUH, such as the angle of pitching, rolling and yawing, the height from the bottom of the water, and the longitude and latitude to calculate the concrete location of the buckets.

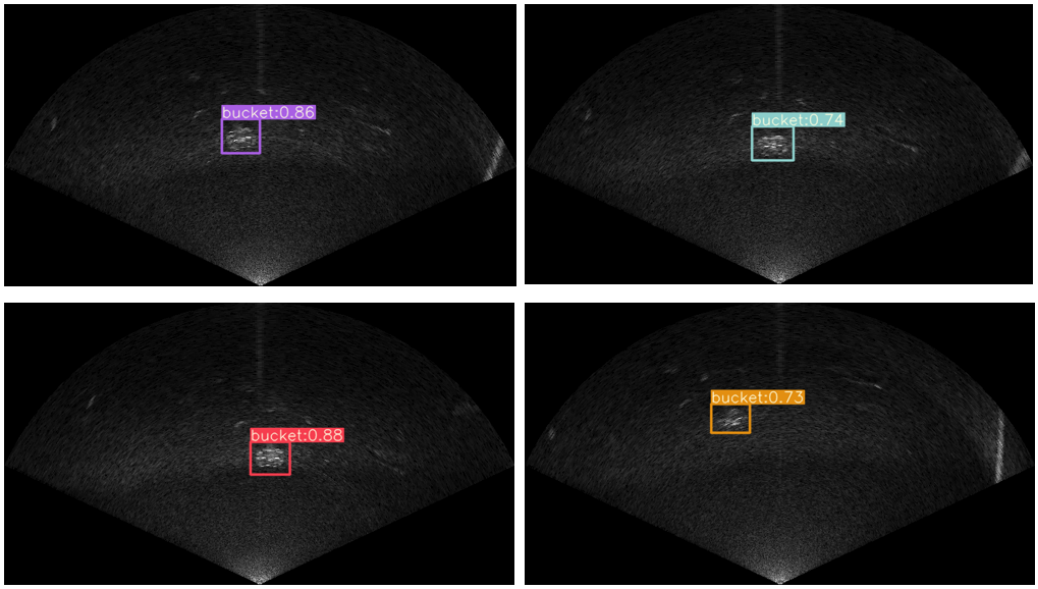


Figure 18. Detection results of the target (bucket).

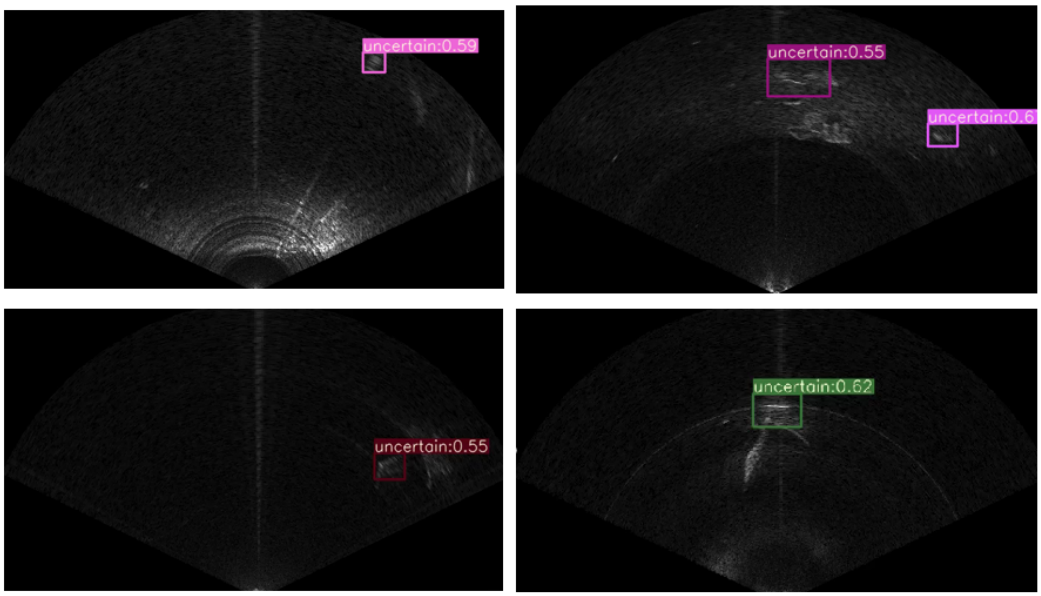


Figure 19. Uncertain detection results.

Figure 20 shows the format of information that the detection module delivered to the main controller of the AUH, which mainly consists of the longitude and latitude of the recognized target and AUH and the category of the target. Then, four recognized targets were pictured in the map using ArcGIS software and were compared to the real locations of the buckets, which were similar to the coordinates marked before.

The results have proven the feasibility and accuracy of the real-time detection system mounted on the AUH and based on the forward-looking sonar imagery. The key techniques of the design of the network are effective, which will have great guiding significance in practice.

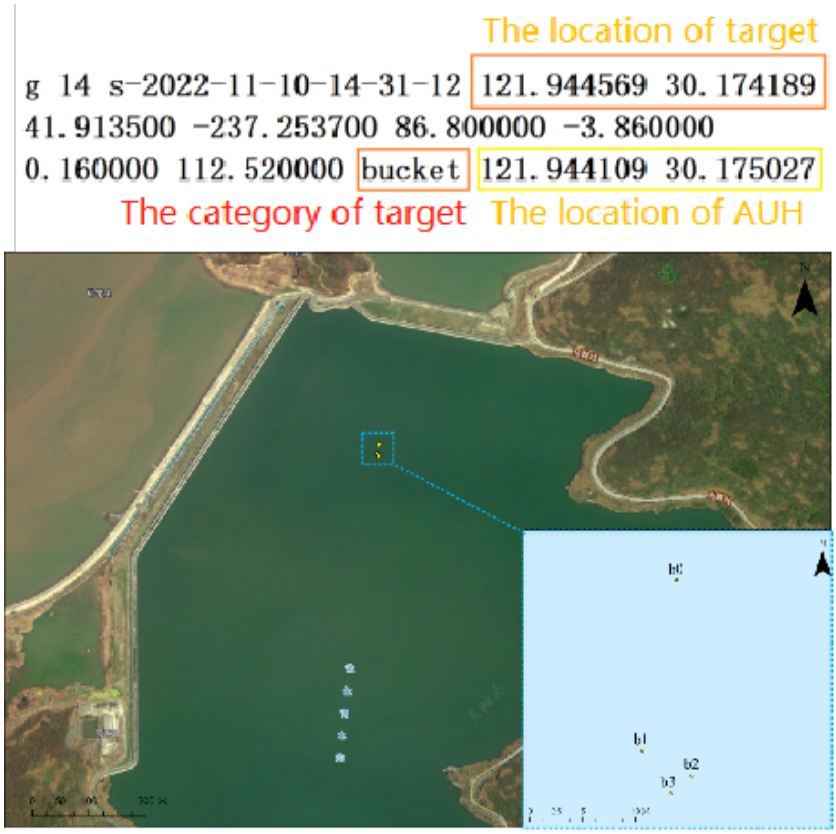


Figure 20. An example of the target message sent to the main controller and a map of the geographical locations of the targets. b0-b3 are the coordinates of the targets given by the detector.

4. Discussion

4.1. Significance of the Proposed Method

1. Theoretically
   The theory of the proposed improved-YOLOv5 is based on many previous works that have received validation. The principles of the attention mechanism have been carefully studied across the literature and have proven that CBAM, CA and GAM are all applicable to sonar-imagery detection and that they can improve the accuracy of most detection models to an extent of about 1–2% on our datasets. Therefore, attention mechanisms can force the models to focus on specific areas of which we want. Furthermore, BiFPN and decoupled heads can boost the mAP by 2–3% on our dataset in a relatively simple way and can achieve a 80.2% high mAP.
   In the meantime, we performed some comparison experiments with other SOTA models, including one-stage and two-stage CNN methods, to prove the advancements of the proposed improved-YOLOv5s. We fed dataset 1 + 2 into the different models and kept the training parameters as similar as possible.
   According to the results shown in Table 8, when attention mechanisms, BiFPN and decoupled heads are added into YOLOv5s, most of them can achieve better results than the other SOTA models—one-stage or two-stage.
2. Systematically
   From the installation of forward-looking sonar to the deployment of the detection model, from the training of an improved network to the validation of efficiency of the detector, and from the communication between the main controller and the detection module to the conversion of the target's location, the research was covered. This real-time AUH-based underwater target detection system using sonar images was designed and test and achieved some significant achievements.
3. Practically

The whole workflow was validated through tank experiments and outdoor tests in open water. A series of experiments give strong support for the facility and efficiency of the designed detection system. The subsea AUH and some simple sensors can achieve the detection goal and detect the desired targets. Moreover, the system provides some thoughts for overcoming the difficulties of underwater target detection. Even with the low-cost sensors and low-quality images, good results can still be achieved by using the stable mounted platform and improved models.

Table 8. Accuracy comparison among different SOTA models.

Methods		mAP (0.5:0.95)
YOLOseries	YOLOv3	0.605
	YOLOv4	0.603
	YOLOv5	0.769
Other One-stage Detectors	RetinaNet	0.753
	SSD	0.780
Two-Stage Detector	Faster-RCNN	0.617
Proposed Methods	+CBAM	**0.77**
	+CA	**0.79**
	+GAM	**0.788**
	+BiFPN	**0.79**
	+decoupled heads	**0.802**

4.2. Limitations of the Proposed Method

1. Theoretically
   Although attention mechanisms can improve mAP, neither the connection between the datasets nor the efficiency of the methods were found, for example, the reason that CBAM only improves by 0.1% mAP when applied to Datasets 1 and 2 but improves by 2% when joining Dataset 3 and why CA performs best in Datasets 1 and 2 but GAM wins in Dataset 3. This information seems to be random, and the reasons that explain these phenomena still need to be studied.
   On the other hand, decoupled heads have introduced more parameters and time to inference, which should be considered carefully, even though it is an efficient way to boost the mAP.
2. Systematically
   Because of the time and cost restrictions of the outdoor experiments, the ablation tests, in reality used to verify the superiority of the improved-YOLOv5 network compared to the raw one, are not finished. The conclusion is qualitative such that the improved network is useful and efficient, but it is not quantitative. In the future, it needs to be proven that in an open water area, when the AUH is on a fixed track and obtains the same amount of sonar images, improved-YOLOv5 can recognize more targets and detect them faster. How to convert longitude and latitude of the targets more precisely through the coordinates of the bounding boxes is also a direction to revise and improve.
3. Practically
   Currently, the proposed method for detecting underwater targets is restricted by the requirements of sonar images. The sonar images of the desired target for detection should be acquired to train the models. However, sometimes it is hard and inconvenient to obtain the desired images; thus, using other techniques such as GAN [42] and diffuse models to generate sonar images with the limited images that already exist without using sonar is a topic that is worthy of discussion. The uncertain targets

cannot be clarified. Thus, aligning other methods and devices to detect them together will make the detection system more practical, for example, using cameras in clean water or magnetometers to detect magnetic targets.

5. Conclusions

A real-time AUH-based underwater detection system, using sonar images, tank experiments, and outdoor tests in open water, was proposed. The main framework of the detector is centered on the improvements of YOLOv5. A series of experiments were conducted and have proved that CBAM, CA, and GAM can improve accuracy by around 1–2% in different datasets, barely gaining parameters and increasing inference time, which is of great significance for future hardware implementation mounted onto AUHs. In addition, by changing YOLO's FPN and PAN neck to BiFPN and by changing the coupled heads to decoupled heads, a 2–3% increment of mAP can be achieved and the models converge faster and more easily.

As publicly known, the key to the widespread application of this real-time target detection system is to improve accuracy while guaranteeing the inference speed and it being lightweight. The proposed methods try to balance this problem well by adding the attention mechanism and by changing FPN to BiFPN without increasing extra time. However, there are still some useful structures such as decoupled heads that can cause problems. Thus, research on model slimming, quantification and distilling should be put forward to keep the system lightweight while improving accuracy.

In conclusion, the research in this paper has great guiding significance in the development of how to solve issues in target detection in complex environments, especially in underwater scenarios. The methods of using AUH and improved detection models can compensate for the deficiencies that are caused by low-quality sonar images captured by budget-friendly sonar and the confusing underwater sound fields. The next step will be to continually adopt the methods mentioned in this paper and perform more online real-time detection sea trials to validate and to compare the performance of different improved models.

Furthermore, optimizing the network is key, for example, trying to introduce another framework such as transformer, which includes another kind of attention mechanism called self-attention to break through the CNN's limitations in capturing global context information. Transformers are more robust to severe occlusions, perturbations and domain shifts compared to CNN frameworks, according to Muzammal [43]. Melting transformer blocks to CNNs can create more context information in the network and can capture more distinguishable feature representations. Some other networks such as GAN and the diffused model will also be used to preprocess and generate more sonar images so as to solve the long-standing problems of poor quality, low resolution and sparsity of datasets, which plague the development of underwater target detection and deep learning.

Author Contributions: Conceptualization, R.C.; Methodology, R.C.; Software, R.C.; Validation, R.C.; Formal analysis, R.C.; Data curation, R.C.; Writing—original draft, R.C.; Project administration, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Access to the data will be considered upon request by the authors.

Acknowledgments: We would like to thank the teams involved in the tank and outdoor experiments that supported this project. We would also like to thank the editor and the anonymous reviewers for their valuable comments and suggestions that greatly improved the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
2. Fukushima, K.; Miyake, S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognit.* **1982**, *15*, 455–469. [CrossRef]
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
4. Chin-Hsing, C.; Jiann-Der, L.; Ming-Chi, L. Classification of underwater signals using wavelet transforms and neural networks. *Math. Comput. Model.* **1998**, *27*, 47–60. [CrossRef]
5. Cortes, C.; Vapnik, V. Support vector machine. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
6. Wan, A.; Dunlap, L.; Ho, D.; Yin, J.; Lee, S.; Jin, H.; Petryk, S.; Bargal, S.A.; Gonzalez, J.E. NBDT: Neural-backed decision trees. *arXiv* **2020**, arXiv:2004.00221.
7. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **1989**, *2*, 396–404.
8. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
10. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
11. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
12. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
16. Wang, J.; Shan, T.; Chandrasekaran, M.; Osedach, T.; Englot, B. Deep learning for detection and tracking of underwater pipelines using multibeam imaging sonar. In Proceedings of the IEEE International Conference on Robotics and Automation Workshop, Montreal, QC, Canada, 20–24 May 2019.
17. Sung, M.; Kim, J.; Lee, M.; Kim, B.; Kim, T.; Kim, J.; Yu, S.C. Realistic sonar image simulation using deep learning for underwater object detection. *Int. J. Control. Autom. Syst.* **2020**, *18*, 523–534. [CrossRef]
18. Lee, S.; Park, B.; Kim, A. Deep learning from shallow dives: Sonar image generation and training for underwater object detection. *arXiv* **2018**, arXiv:1810.07990.
19. Chen, R.; Zhan, S.; Chen, Y. Underwater Target Detection Algorithm Based on YOLO and Swin Transformer for Sonar Images. In Proceedings of the OCEANS 2022, Hampton Roads, VA, USA, 17–20 October 2022; pp. 1–7.
20. Dobeck, G.J. Algorithm fusion for the detection and classification of sea mines in the very shallow water region using side-scan sonar imagery. In *Detection and Remediation Technologies for Mines and Minelike Targets V*; SPIE: Bellingham, WA, USA, 2000; Volume 4038, pp. 348–361.
21. Jing, Y.; Ren, Y.; Liu, Y.; Wang, D.; Yu, L. Automatic extraction of damaged houses by earthquake based on improved YOLOv5: A case study in Yangbi. *Remote Sens.* **2022**, *14*, 382. [CrossRef]
22. Panboonyuen, T.; Thongbai, S.; Wongweeranimit, W.; Santitamnont, P.; Suphan, K.; Charoenphon, C. Object detection of road assets using transformer-based YOLOX with feature pyramid decoder on thai highway panorama. *Information* **2022**, *13*, 5. [CrossRef]
23. Xu, C.; Wang, X.; Yang, Y. Attention-YOLO: YOLO Object Detection Algorithm with Attention Mechanism. *Comput. Eng. Appl.* **2019**, *55*, 12.
24. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight underwater object detection based on yolo v4 and multi-scale attentional feature fusion. *Remote Sens.* **2021**, *13*, 4706. [CrossRef]
25. Kong, W.; Hong, J.; Jia, M.; Yao, J.; Cong, W.; Hu, H.; Zhang, H. YOLOv3-DPFIN: A dual-path feature fusion neural network for robust real-time sonar target detection. *IEEE Sens. J.* **2019**, *20*, 3745–3756. [CrossRef]
26. Topple, J.M.; Fawcett, J.A. MiNet: Efficient deep learning automatic target recognition for small autonomous vehicles. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1014–1018. [CrossRef]
27. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

28. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
29. Wang, Z.; Liu, X.; Huang, H.; Chen, Y. Development of an autonomous underwater helicopter with high maneuverability. *Appl. Sci.* **2019**, *9*, 4072. [CrossRef]
30. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [CrossRef]
31. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
32. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef] [PubMed]
33. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.
34. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
35. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
36. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
37. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
38. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
39. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
40. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
41. Zhou, Y.; Chen, S.; Wu, K.; Ning, M.; Chen, H.; Zhang, P. SCTD1. 0: Sonar common target detection dataset. *Comput. Sci.* **2021**, *48*, 334–339.
42. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
43. Naseer, M.M.; Ranasinghe, K.; Khan, S.H.; Hayat, M.; Shahbaz Khan, F.; Yang, M.H. Intriguing properties of vision transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23296–23308.