Journal of
Marine Science and Engineering

MDPI

Article

Research on the Visual Perception of Ship Engine Rooms Based on Deep Learning

Yongkang Wang, Jundong Zhang *, Jinting Zhu, Yuequn Ge and Guanyu Zhai

College of Marine Engineering, Dalian Maritime University, Dalian 116026, China; wyk_9825@dlmu.edu.cn (Y.W.)
* Correspondence: zhjundong@dlmu.edu.cn

Abstract: In the intelligent engine room, the visual perception of ship engine room equipment is the premise of defect identification and the replacement of manual operation. This paper improves YOLOv5 for the problems of mutual occlusion of cabin equipment, an unbalanced number of different categories, and a large proportion of small targets. First, a coordinate attention (CA) mechanism is introduced into the backbone-extraction network to improve the ability of the network to extract main features. Secondly, this paper improves the neck network so that the network can learn a relatively important resolution for feature-fusion and enrich the semantic information between different layers. At the same time, this paper uses the Swin transformer as the prediction head (SPH). This enables the network to establish global connections in complex environments, which can improve detection accuracy. In order to solve the problem of cabin equipment covering each other, this paper replaces the original non-maxima suppression (NMS) with Soft-NMS. Finally, this paper uses the K-means algorithm based on the genetic algorithm to cluster new anchor boxes to match the dataset better. This paper is evaluated on the laboratory's engine room equipment dataset (EMER) and the public dataset PASCAL VOC. Compared with YOLOv5m, the mAP of CBS-YOLOv5m increased by 3.34% and 1.8%, respectively.

Keywords: engine room equipment detection; EMER dataset; coordinate attention; Swin transformer; soft-NMS

check for updates

Citation: Wang, Y.; Zhang, J.; Zhu, J.; Ge, Y.; Zhai, G. Research on the Visual Perception of Ship Engine Rooms Based on Deep Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 1450. https://doi.org/10.3390/jmse11071450

Academic Editor: Leszek Chybowski

Received: 24 June 2023
Revised: 15 July 2023
Accepted: 18 July 2023
Published: 20 July 2023

1. Introduction

With the development of science and technology, ships are made to be large-scale, modernized and intelligent. At the same time, the number of crew members is simultaneously being reduced, which makes the intelligent engine room the current research hotspot of intelligent ships. At present, the unmanned and intelligent engine room mainly depends on the engine room's intelligent monitoring and alarm system. The engine room monitoring and alarm system can monitor the operating status of each system of the ship in real-time to ensure the regular operation of the ship. When a fault occurs, the system will send out an audible and visual alarm and record the operating and system status data at the same time, which is convenient for the engineer when carrying out maintenance. However, nowadays, the monitoring system still needs improvement. For the appearance defects of equipment, leaking, and valve operation, engineers still need to go to the engine room to conduct inspections—for example, if the engine room is unoccupied and a pipe suddenly leaks. If not found immediately, the consequences may be fatal.

If the ship's equipment defects during navigation cannot be found in time, this will result in pipeline leakage, which will cause unnecessary losses; if the ship is in stormy weather, casualties are likely to occur when engine room personnel operate valves. Therefore, visual sensors are used to identify the appearance status of the equipment, and the information is integrated into the monitoring system so that faults and defects can be found early, thereby helping engineers to deal with hidden dangers in advance. Similarly, robots, which can operate instead of the engineer, utilize visual sensors to identify the valves in the engine room. Realizing these ideas mainly depends on technologies such as computer vision, and applying computer vision and other technologies to the engine room plays an important role in realizing the intelligent engine room.

However, few visual perception technologies are currently used in the engine room. The primary purpose of this paper is to use computer vision technology to identify ship engine room equipment, thereby replacing the eyes of the engineer to identify the equipment when no one there is on duty. At the same time, it provides potential guidance for the subsequent inspection of the appearance of equipment and operations, in place of engineers. Although the convolutional neural network has achieved good results in the stability and accuracy of object detection, the equipment monitoring task in the actual engine room still faces various difficulties and challenges, including the following aspects:

1. There are no datasets available for ship engine rooms.
2. The number of meters and valves in the equipment accounts for a large proportion. However, the proportion of other equipment is tiny. Valves and meters are generally small and medium objects in images. Therefore, there are problems of class imbalance and the dense distribution of small targets in the cabin.
3. The scale of equipment is enormous, from large diesel engines to tiny valves. Moreover, the environment of the engine room is complex, the arrangement of equipment is compact, and the distribution of pipelines is dense, resulting in many concealment problems.

In response to the above challenges, this paper proposes an improved equipment-detection model based on YOLOv5m to achieve the detection accuracy requirements in engine room equipment. These are the paper's contributions:

1. Use the laboratory 3D virtual engine room team to select the obtained engine room pictures and establish the ship engine room equipment datasets (EMER), in which the equipment categories include engine, separator, cooler, reservoir, pump, valve, and meter.
2. This paper uses the improved CSPDarknet53 network base on coordinate attention (C3CA) to improve the ability of the backbone network to extract important features to improve the detection ability. The neck network uses the Weighted Bidirectional Feature Pyramid structure of Vertical and Horizontal connections (VH-BiFPN) that we designed, which enriches the semantic information of different resolutions and achieves the goal of improving the detection ability. We design the Swin transformer detection head to obtain the global information of the feature map and the connection of the context to enhance the model's ability to detect small targets.
3. According to the engine room's distribution characteristics, soft-non-maxima suppression (Soft-NMS) reduces the mutual masking problem in the engine room, reduces missed detection in the engine room, and enhances the recall and accuracy of equipment detection in the engine room.

The remainder of the paper is structured as follows. Section 2 addresses relevant object-detection research; Section 3 introduces the improved YOLOv5m based on CA, VH-BiFPN and the Swin transformer (CBS-YOLOv5m) detection model. In the fourth section, the model is verified on the PASCAL VOC dataset, and the EMER dataset and ablation experiments are performed. The fifth section includes the conclusion and discussion.

2. Related Work

2.1. Data Augmentation

Data augmentation can prevent model overfitting and improve the model's generalization ability. Optical distortion and geometric distortion are frequently used data augmentation methods. For optical distortion, we make adjustments to the hue of the image. We perform random scaling, cropping, translation, and rotation for geometric distortion. In addition to the aforementioned global pixel-enhancement techniques, there are also unique enhancement methods. Some researchers have proposed methods for data augmentation by combining multiple images, namely Mixup [1] and Mosaic [2]. Mixup randomly selects two samples from the training images and performs a weighted summation, where the labels of the samples are also weighted for summation. Mosaic randomly selects four images from the training images, which significantly enriches the background of the detected object.

In this article, not only traditional methods are used for enhancement, but also Mixup and Mosaic methods are used.

2.2. Object Detection

Object detection is divided into traditional object t detection and deep learning object detection. Traditional object-detection algorithms often take a lot of time and manual processing of datasets, such as Histograms of Oriented Gradients [3] and Scale Invariant Feature Transform [4]. In contrast, data-driven deep learning detection algorithms can automatically extract features from data, which greatly reduces the burden of traditional design and improves detection accuracy.

Currently, there are two methods for object detection: two-stage object-detection methods and one-stage object-detection methods. The two-stage detection method generates candidate regions on the image first, extracts feature through a convolutional neural network, and then performs classification and regression one by one. Region-CNN (R-CNN) [5] is the first two-stage object-detection algorithm. Although it reduces window redundancy and algorithm time complexity, it requires a fixed-size input image. In order to overcome this shortcoming, Ref. [6] proposed the Spatial Pyramid Pooling Network (SPPNet) network, which can extract features in any area without a specific image size input and reduce the amount of calculation. Ref. [7] combined the ideas of R-CNN and SPPNet, and Fast R-CNN was proposed. It improves detection speed and accuracy by using a special case of SPP, which is the region of interest pooling layer (ROI). Ref. [8] proposed Faster R-CNN on the basis of Fast R-CNN, which uses a fully convolutional network to generate candidate regions on the feature map so that it can improve the quality of candidate regions and thus greatly improve the speed. However, the detection speed of the above two-stage target algorithm is relatively slow.

One-stage object-detection methods directly predict detection boxes and class probabilities. For example, Single-Shot MultiBox Detector (SSD) [9] predicted using multi-scale feature maps according to the characteristics of semantic information at different levels for the first time. However, it is prone to the problem of imbalance between foreground and background classes. In order to avoid this problem, RetinaNet [10] designed the focal loss function and used the Feature Pyramid Network (FPN) [11] to improve the detection performance of different scale targets. Another representative algorithm is You Only Look Once (YOLO) and its variants. YOLO [12] used a single network, which greatly improves the speed, but its accuracy and recall are low. YOLOv2 [13] combined a variety of new modules to improve YOLO and introduced an anchor frame mechanism similar to Faster R-CNN, which improved the accuracy and recall. Further, YOLOv3 [14] introduced a multi-scale framework, multi-scale feature-fusion and residual structure, which improved the detection accuracy of small targets. YOLOv4 [2] employed a cross-stage part based on Darknet-53, reducing the calculation amount. YOLOv7 [15] utilized E-ELAN as the backbone network to extract image features based on ELAN. Then, an auxiliary detection head was designed to improve the detection accuracy. In addition, structural re-parameterization was introduced to improve detection speed. At present, YOLOv5 is the most stable generation of the YOLO series, and it has excellent performance on different datasets.

In the detection of ship equipment, the scale difference between different pieces of equipment is too large. There are huge differences in the proportions of different devices in the same image, and the proportions of the same device in different images are different. Figure 1a–c shows large objects, medium objects, and small objects, respectively. Figure 1d is a multi-scale object. Currently, different feature layers are used to predict different scale targets. For example, YOLOv5 used 80 × 80 feature maps to predict small targets, 40 × 40 feature maps to predict medium targets, and 20 × 20 to predict large targets. Ref. [16] improved various devices' detection speed and recognition accuracy by introducing structural re-parameterization and the Neighbor Erasing and Transferring Mechanism (NETM) in RetinaNet. Ref. [17] improved the detection speed of the device by using pruning operations on the YOLO algorithm, but their detection accuracy for meters and valves is low; that is, the detection accuracy for small targets is low.

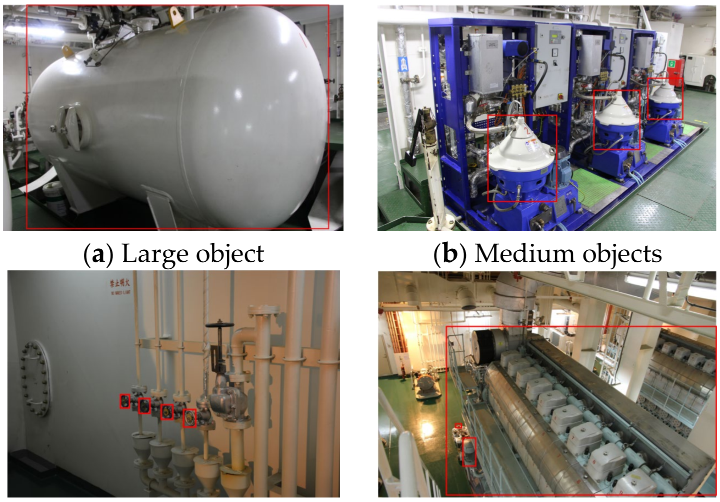

(**a**) Large object (**b**) Medium objects
(**c**) Small objects (**d**) Multi-scale objects

Figure 1. Multi-scale example of engine room equipment. In (**a**), equipment accounted for almost 100%. In (**b**), equipment accounts for about 10%. (**c**) shows a small target accounting for 1%, and (**d**) shows the multi-scale variation between devices.

2.3. Attention Mechanism

The attentions applied in convolutional neural networks (CNNs) include the spatial transformation network [18] that realizes spatial attention, the Squeeze-and-Excitation Network [19] that realizes channel attention, and the convolutional block attention module that realizes channel attention and spatial attention (CBAM) [20]. These lightweight attention modules can be directly applied to a CNN to improve the model's extraction of crucial information. The latest research direction of the attention mechanism is self-attention. It was inspired by the Natural Language Processing (NLP) field, using the self-attention layer to replace the convolution for image-processing tasks. Refs. [21,22] used pure self-attention deep networks and obtained state-of-the-art (SOTA) results in the image field, but the amount of calculation is enormous. This shows that self-attention models have great potential for detection performance. The Swin transformer [23] uses a self-attention method with shifted windows, achieving many SOTA results in image processing, which shows the superiority of self-attention-based attention mechanisms in image tasks. Therefore, combining an attention mechanism and convolution can significantly improve the model's performance. For example, Ref. [24] combined the attention mechanism of the Swin transformer and the Normalization-based Attention Module (NAM) [25] in the convolutional neural network and achieved good results in remote sensing images.

In response to the difficulties in related work and cabin equipment image detection, an improved YOLOv5m cabin equipment-detection model is proposed, which is named CBS-YOLOv5m.

3. Methodology

In this section, we first provide an overview of the structure of YOLOv5 and discuss its application's shortcomings in the engine room. Then, we introduce structures such as the Swin transformer, coordinate attention (CA), Soft-NMS, and improved Weighted Bidirectional Feature Pyramid (BiFPN) to optimize the dense scene in the engine room.

3.1. YOLOv5

Up to now, YOLOv5 is the most stable version of the YOLO series in different datasets. According to different network depths and widths, it can be divided into five types of networks: n, s, m, l, and x. Due to the limited number of datasets used in this paper, YOLOv5m is adopted as the benchmark network. Figure 2 shows the original YOLOv5 network structure.

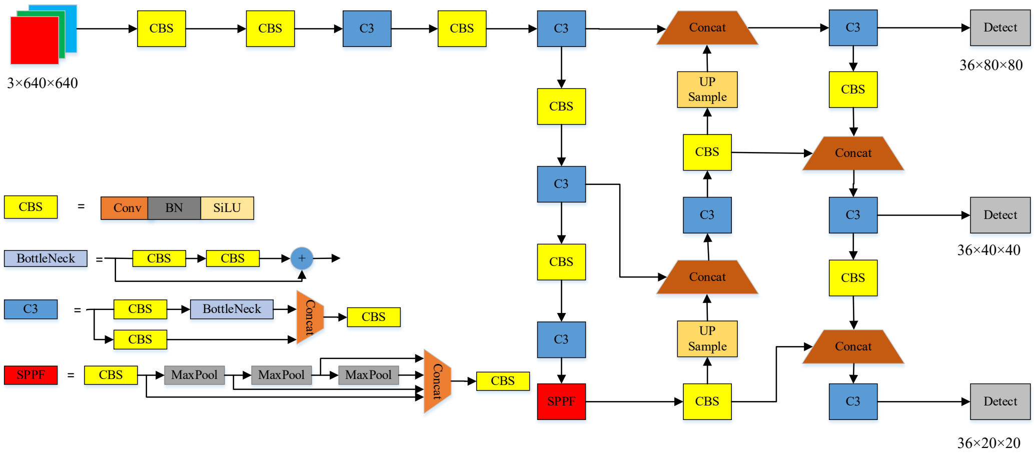


Figure 2. Original YOLOv5 Network Structure.

In the input, YOLOv5 uses the enhanced function of Mosaic to randomly combine four pictures in the dataset into a 640 × 640 picture, which can enhance the complexity of the dataset. The first part of this structure is the backbone network. It consists of BottleNeckCSP and Spatial Pyramid Pooling (SPP) modules. Compared with ResNet [26], BottleNeckCSP reduces a 1 × 1 convolution kernel calculation to reduce computational complexity and extract depth information from features more effectively. The SPP module can increase the network's accepted domain by pooling in different ranges. The second part is the neck network combines the operations of the Path Aggregation Network (PANet) and FPN. High-level feature-dense localization employs the PANet framework. Meanwhile, FPN provides underlying semantic features through up-sampling. By combining the two structures, semantic features of different scales can be better integrated, which can enhance the detection performance. Finally, the head network classifies feature maps of different scales. The output includes category probability, confidence score, and bounding box information.

YOLOv5 has an excellent detection performance, but it also has certain limitations:

1. It is mainly applicable to COCO datasets but cannot necessarily be applied to datasets in certain specific scenarios.
2. The PANet structure focuses less on information between non-adjacent levels, resulting in a decrease in information during each fusion process.
3. For dense scenarios, the NMS processing mechanism is simple and rough, which can easily lead to missed detections and lower recall rates.
4. It lacks the ability to capture global and contextual information and cannot effectively utilize the positional and spatial information of feature maps.

3.2. Proposed Structural Improvements

We propose to improve the original YOLOv5 model for detecting small targets in the engine room. To further improve accuracy without significantly increasing model complexity, we introduce several improvements and propose a network structure, as shown in Figure 3.

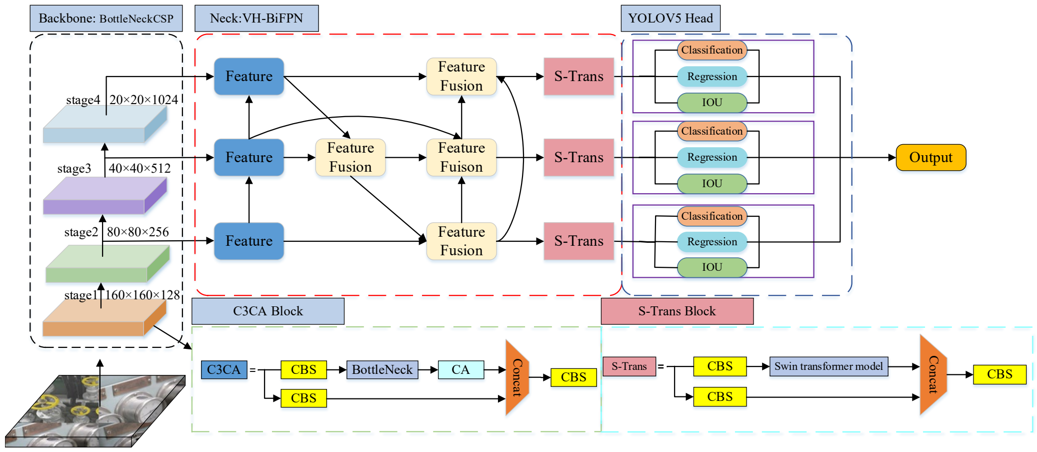


Figure 3. Network structure diagram.

3.2.1. Improved Backbone

In order to prevent the loss of important information in the extracted features, we add the coordinate attention structure [27] to the backbone network, which obtains the importance of each coordinate on the image by averaging pooling in different directions of the image. This allows more focus on important features when extracting features, thereby overcoming the loss of important features. The coordinate attention mechanism is a lightweight attention mechanism that does not increase computational overhead. The C3 module is crucial for extracting feature information in the backbone network. Therefore, this paper adopts a design that combines C3 and CA modules to enhance the ability to extract image features.

The CA mechanism performs average pooling in different directions of the image, then encodes through operations such as convolution, and finally finds the importance of each coordinate and then multiplies it with the original feature so that it focuses on the important feature. The CA module is shown in Figure 4.

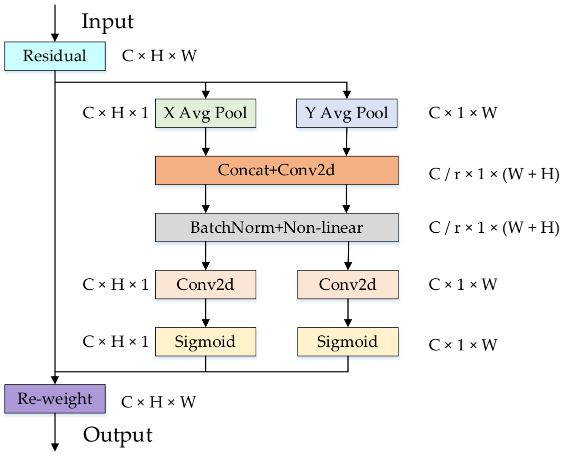


Figure 4. CA Module.

This article compares commonly used attention mechanisms and verifies the ability of the model through mAP, a common indicator of object detection, which reflects the coverage ability and recognition accuracy of the classifier for positive samples. The mAP used in this paper is calculated when the IOU threshold is 0.5. In Table 1, CA, CBAM, and SE are compared. The experimental results show that CA performs better in the engine room dataset and can effectively extract features in the engine room dataset. Therefore, this paper adopts the method of CA binding to C3.

Table 1. Comparison of different attention mechanisms.

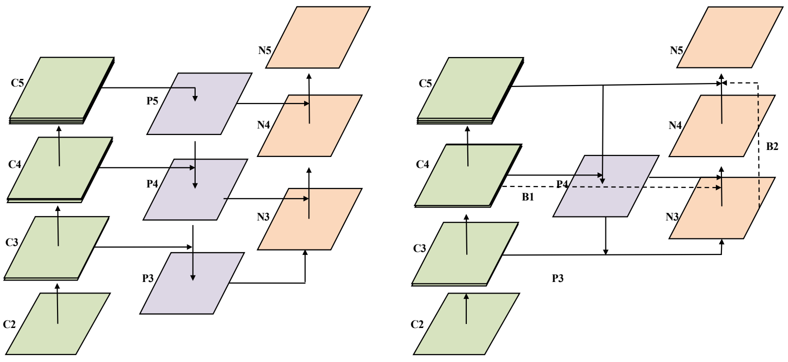
Model	mAP (%)
YOLOv5m	83.73
YOLOv5m+SE	84.53 (+0.8)
YOLOv5m+CBAM	84.33 (+0.6)
YOLOv5m+CA	84.71 (+0.98)

The article combines the C3 module with the CA module and adds the CA module to the BottleNeck module in C3 to better extract important image features and locate targets more accurately. The designed C3CA module is shown in Figure 3.

3.2.2. Improved PANet Structure

The Neck network is to fuse the features extracted by the backbone network so that the head can detect targets with different proportions according to different resolutions. It usually consists of top-down and bottom-up paths to achieve the purpose of passing shallow semantic information to deep layers [11] and deep semantic information to shallow layers. This avoids the problem of the loss of feature information as the convolutional network deepens. However, this approach relies on the aggregation of adjacent layer features and pays less attention to the information exchange of non-adjacent layers. So, with each aggregation, the spatial information of non-adjacent layers is continuously reduced.

The research in [28] shows that the upper and lower parts of the middle layer, as shown in Figure 5a, have little effect on the overall feature-fusion effect, so they can be removed, which can reduce the computational overhead. As shown in Figure 5b, we increase the cross-layer connection (B1) in the horizontal direction, which can solve the problem of missing original feature information caused by the information-fusion process from the C-N part. In order to solve the problem of less information exchange in non-adjacent layers, this paper adds a vertical cross-layer connection (B2), which can enrich semantic information to the greatest extent.



(**a**) PANet structure (**b**) Improved PANet structure

Figure 5. Improved Neck comparison.

In the YOLOv5m feature-fusion stage, the channels are directly spliced, which can quickly fuse information at different levels [29]. Inspired by [28], the contributions of feature channels at different resolutions should be different [30]. Therefore, we assign learnable weights to different resolutions and find the optimal weight ratio through continuous updating.

The formula for rapid normalization is expressed as:

$$O = \sum_{i} \frac{w_i}{\epsilon + \sum_{j} w_j} \cdot I_i \tag{1}$$

The learnable weights are thus expressed as follows:

$$P_i^{td} = \mathrm{Conv}\left(\frac{w_1 P_i^{in} + w_2 \mathrm{Resize}\left(P_{i+1}^{td}\right)}{w_1 + w_2 + \varepsilon}\right) \tag{2}$$

$$P_i^{\mathrm{out}} = \mathrm{Conv}\left(\frac{w_1' P_i^{\mathrm{in}} + w_2' \mathrm{Resize}\left(P_i^{\mathrm{td}}\right) + w_3' \mathrm{Resize}\left(P_{i-1}^{\mathrm{out}}\right)}{w_1' + w_2' + w_3' + \varepsilon}\right) \tag{3}$$

where w_i is a learnable weight that represents the importance of input features, and ε is set to 0.001 to prevent the denominator from being set to 0. P_i^{td} represents the intermediate features of level i on the top-down path, and P_i^{out} represents the output features at the level of i on the top-down path. All features are established using this similar method.

3.2.3. Improved Head Structure

This section is focused on the problem of the loss of detailed information in the complex scene YOLOv5 feature-extraction process in the cabin image. Inspired by the vision transformer, this paper uses the global information capture capability of the transformer encoder block to improve network-detection capabilities. However, if the transformer is in the visual field, the amount of calculation will be greatly increased. At the same time, Microsoft Research Asia proposed the Swin transformer, which significantly reduces the amount of calculation compared with the vision transformer. The computational complexity of the two is as follows:

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C \tag{4}$$

$$\Omega(\mathrm{W-MSA}) = 4hwC^2 + 2M^2hwC \tag{5}$$

The former is the computational complexity of the vision transformer, which is the quadratic curve of hw. The latter is the computational complexity of the Swin transformer, and M is a constant 8, which is a linear function of hw. This dramatically reduces the computational complexity.

The Swin transformer module is shown in Figure 6. It consists of Layer Normalization (LN), Window-based Multi-head Self-Attention (W-MSA), Shifted Window-based Multi-head Self-Attention (SW-MSA) and two layers of Multilayer Perceptron (MLP). After attention and MLP, there is a layer of dropout [23]. Since the window-based self-attention connections are localized in the partitioned windows, it limits the capability of the model. The multi-head self-attention based on the shifted window can connect across windows, so the two modules are used in pairs in Swin transformer to indirectly achieve global information comparison.

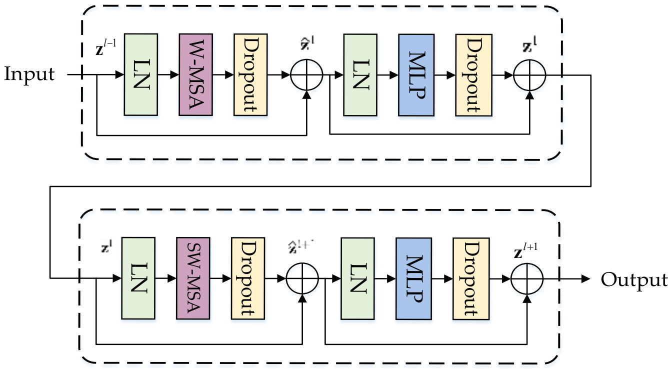


Figure 6. Swin transformer module.

In the Swin transformer, multi-head self-attention is used for computation in the window to achieve the ability to capture information globally and obtain contextual relationships. This is shown in Figure 7.

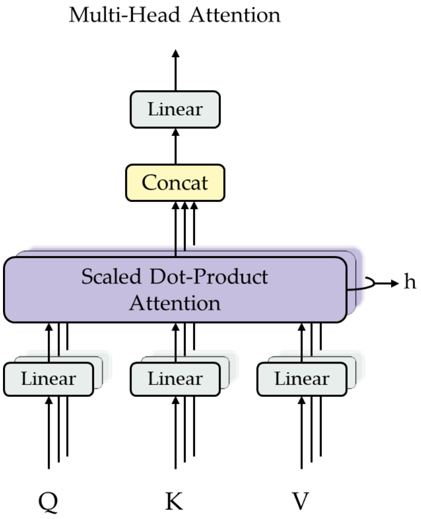


Figure 7. Schematic diagram of multi-head self-attention.

When calculating self-attention, relative position bias encoding is used for the calculation, as follows:

$$Attention(Q,K,V) = SoftMax(QK^T/\sqrt{d} + B)V \tag{6}$$

$$(S)MSA(Q,K,V) = Concat(head_1,\ldots,head_h)W^O \tag{7}$$

where $head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i^V)$; $Q,K,V \in \mathbb{R}^{M^2\times d}$ are query, key, and value matrices, respectively; d is the dimension of the query/key. Since the relative positions of each axis are within the range of $[-M+1, M-1]$, we parameterize the smaller size of the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M-1)}$, and the value of B is taken from $\hat{B}$ [23].

The Swin transformer achieves the effect of global attention in another way—by connecting two different modules in series and reducing the computational complexity—but the computational complexity is still enormous. Therefore, we only apply it to the head, forming the Transformer Prediction Head (SPH). This is due to the low resolution of the feature maps in the head network. Using SPH on low-resolution feature maps can reduce computational complexity. The S-Trans Block after replacing the Bottleneck with the Swin transformer model is shown in Figure 3.

3.2.4. Soft-NMS

According to research on the dataset, there are various pieces of equipment covering each other in the ship engine room, such as similar equipment obstacles, different types of equipment blocking each other, and non-equipment obstacles. Taking the first case as an example, the real boxes of two adjacent devices will generate a series of detection boxes, which obtain different confidence levels compared with the real boxes. In the NMS processing mechanism, these boxes are sorted to select the bounding boxes with the highest confidence, and the rest are deleted. At this time, the detection box generated by similar devices adjacent to it may be mistakenly deleted, resulting in the loss of the detection box of the neighboring devices. As shown in Figure 8, the two separators have the same characteristics. When the two separators overlap, for the separator (S2), the detection frame B1 of the separator (S1) will obtain the highest confidence level of the two separators. Therefore, the real detection frame B2 of the separator (S2) will be deleted, which will lead to the missed detection of the separator (S2).

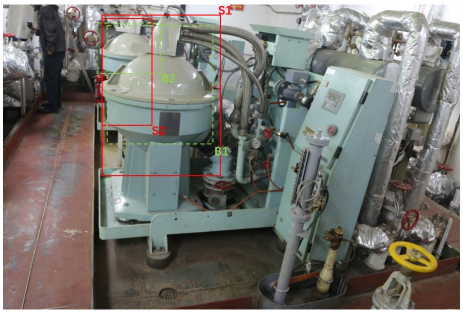
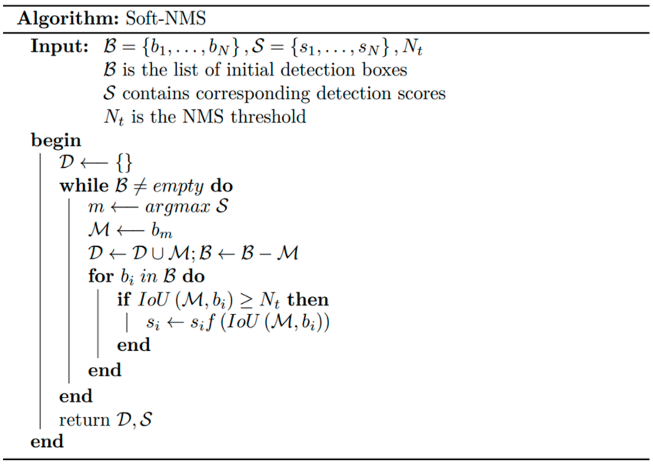


Figure 8. Shows an example where the equipment in the engine room is obscured.

To avoid this omission, we change the original NMS [31] of YOLOv5m to Soft-NMS [32]. That is, boxes with low confidence are not deleted directly but have reduced confidence and are retained. The pseudo-code is shown in Figure 9.



```
Algorithm: Soft-NMS
Input: $\mathcal{B} = \{b_1, \ldots, b_N\}, \mathcal{S} = \{s_1, \ldots, s_N\}, N_t$
       $\mathcal{B}$ is the list of initial detection boxes
       $\mathcal{S}$ contains corresponding detection scores
       $N_t$ is the NMS threshold
begin
    $\mathcal{D} \longleftarrow \{\}$
    while $\mathcal{B} \neq empty$ do
        $m \longleftarrow argmax\ \mathcal{S}$
        $\mathcal{M} \longleftarrow b_m$
        $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{M}; \mathcal{B} \leftarrow \mathcal{B} - \mathcal{M}$
        for $b_i$ in $\mathcal{B}$ do
            if $IoU(\mathcal{M}, b_i) \geq N_t$ then
                $s_i \leftarrow s_i f(IoU(\mathcal{M}, b_i))$
            end
        end
    end
    return $\mathcal{D}, \mathcal{S}$
end
```

Figure 9. The pseudo-code of Soft-NMS.

In this paper, the Soft-NMS method is used to replace the NMS method, and the low confidence score is not directly set to 0. Instead, the confidence score is reduced so that it has the opportunity to participate in the selection of candidate-detection regions, which can reduce the occurrence of missed detection. The Soft-NMS method is calculated as follows:

$$s_i = \begin{cases} s_i, & \mathrm{iou}(\mathcal{M}, b_i) < N_t \\ s_i e^{-\frac{\mathrm{iou}(\mathcal{M}, b_i)^2}{\delta}}, & \mathrm{iou}(\mathcal{M}, b_i) \geq N_t \end{cases} \tag{8}$$

The above function multiplies the corresponding confidence level that is higher than the IOU threshold by a Gaussian function, so the detection far away from $\mathcal{M}$ will not be affected. In contrast, detection boxes that are very close will be assigned a greater penalty.

3.2.5. Anchor Box Clustering

In the target-detection task, the model not only needs to learn the category of the target but also needs to learn the position and size of the target. However, there are objects with different aspect ratios in each image. This makes it more difficult for the model to learn the object's shape. Therefore, the prior box mechanism proposed in [8] divides the space of objects with different scales and aspect ratios into several. It is applied to SSD, YOLOv3, RetinaNet, etc. In YOLOv5, the K-means algorithm based on the genetic algorithm is used to cluster the training set to obtain nine prior boxes as the initial anchor boxes. Because the K-means algorithm easily falls into the local optimal solution, the genetic algorithm applies the "survival of the fittest" principle to make the solution move in a good direction. It can seek the optimal solution globally. Therefore, the two are combined to improve the quality of clustering.

The default anchor box of YOLOv5 is obtained by clustering the COCO dataset, so this paper uses the K-means algorithm based on the genetic algorithm to re-cluster the dataset. In the PASCAL VOC dataset, the new anchor boxes are: 80 × 80 anchor boxes are [[30, 52], [78, 63], [63, 138]], 40 × 40 anchor boxes are [[146, 125], [120, 224], [198, 303]], 20 × 20 anchor boxes are [[389, 207], [324, 411], [536, 394]]. In the EMER dataset, the new anchor boxes are: 80 × 80 anchor boxes are [[9, 9], [13, 14], [19, 19]], 40 × 40 anchor boxes are [[26, 26], [37, 35], [50, 51]], 20 × 20 anchor boxes are [[76, 79], [107, 152], [234, 231]].

4. Experiments

In this part, the effectiveness of the improved YOLOv5 model is verified on the EMER dataset. In Section 4.1, the paper introduces the dataset construction process, as well as the equipment used in the experiment. Finally, in Section 4.2, the paper validates the model through the public dataset PASCAL VOC2012+2007 and EMER dataset and makes an intuitive comparison of some detection results of the EMER dataset. The schematic diagram of engine room equipment-detection training is shown in Figure 10.

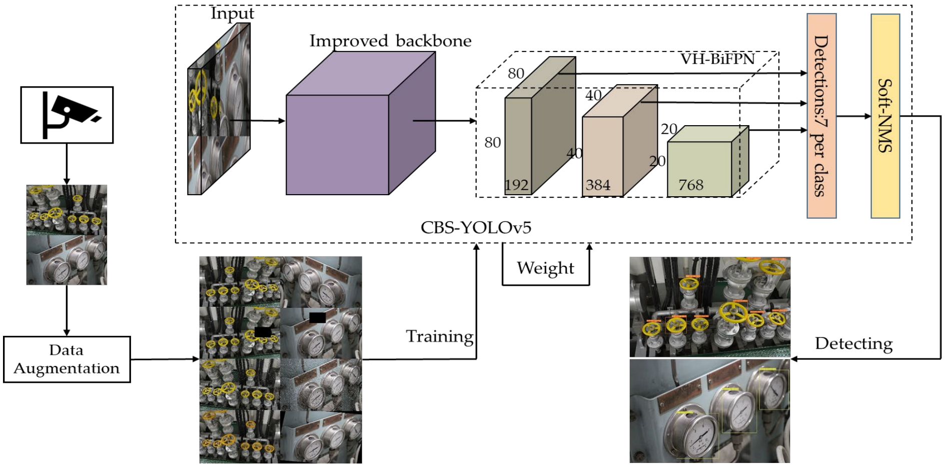
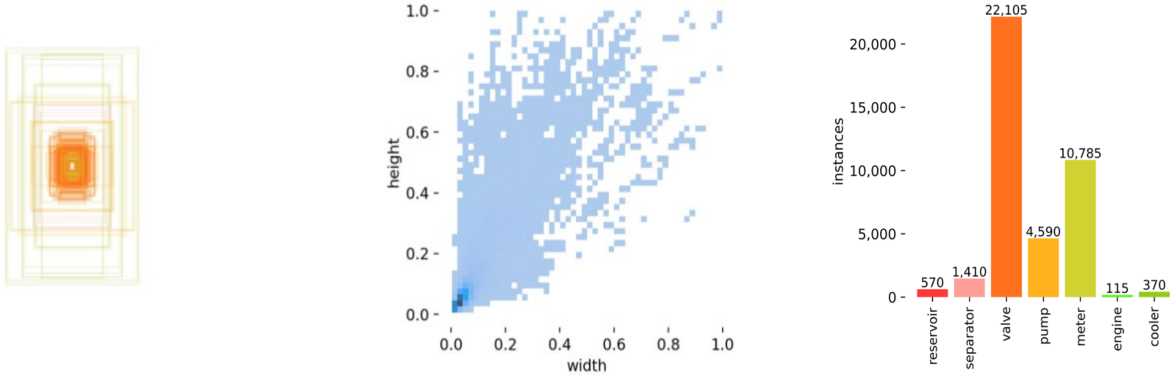


Figure 10. Training and detecting process of engine room equipment.

4.1. EMER Dataset

With the resources of the laboratory team, we screened 1725 images from 4 different ship types. Most of the images were taken with Canon digital cameras, while others were captured via engine room surveillance. Due to some problems, such as different shooting angles, complex cabin environment, changes in cabin lighting, and the unbalanced number of different devices, in order to enhance the generalization ability of the model, we used Gaussian noise, mirroring, rotation, translation, color change, cutout, etc., to expand the dataset. We expanded the original dataset from 1725 to 8625, including 7 types of instruments such as cooler, engine, meter, pump, reservoir, separator, and valve. The ratio of the training set, verification set, and test set is 7:2:1.

Then we analyzed the dataset, as shown in Figure 11: (a) shows the distribution of the position and shape of the label box; (b) represents the normalized target size. It can be seen from the figure that the size of the target is mainly concentrated at 0~0.1, so most of the small targets in the dataset; (c) represents the number of labeled boxes for various types of equipment, of which the number of valves and meters accounts for about 82%. Engines, coolers, and reservoirs account for only 3%. The number of large targets is small, but high detection accuracy can still be obtained after data enhancement. This is because the characteristics of large objects are more obvious. Therefore, this paper mainly focuses on how to improve the detection ability of small targets.



(**a**) (**b**) (**c**)

Figure 11. EMER dataset analysis. (**a**) Anchor box shape distribution. (**b**) Target normalized size. (**c**) Quantity of each type of equipment.

The equipment used in the experiment is shown in Table 2. The operating system is Ubuntu 20.04.1, the graphics processor is NVIDIA GeForce RTX 3090, the CPU is 24-core AMD 3960X, the memory is 64G, the Integrated development environment (IDE) is Visual Studio, the model is built with the programming language Python 3.8.10 and the deep learning framework PyTorch 1.10.0, and CUDA11.3 is used for acceleration during training.

Table 2. Experimental platform.

Configuration	Specification
Operating System	Ubuntu 20.04.1
GPU	NVIDIA GeForce RTX 3090
CPU	AMD 3960X 24-core
RAM	64 G
IDE	Visual Studio
Framework	PyTorch-1.10.0
Toolkit	CUDA 11.3

4.2. Model Validation

4.2.1. Criteria

In object detection, representing these detection boxes as true objects and false objects produces four kinds of predictions: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). If the original sample is a positive sample and the model prediction result is also a positive sample, then it is TP. Otherwise, it is marked as FP. False negative refers to wrongly predicting positive samples as negative samples, and true negative is the opposite. Mean Average Precision (mAP) is a commonly used indicator for evaluating the performance of object-detection models, which is related to recall and accuracy. Usually, Formulas (9) and (10) are used to express precision and recall.

$$Pr = \frac{TP}{FP + TP} \tag{9}$$

$$\mathrm{Re} = \frac{TP}{FN + TP} \tag{10}$$

In the formula, Pr is the accuracy and Re is the recall. Accuracy and recall are mutually related, and the two measures are usually combined to obtain a better evaluation of the model's capabilities. The recall is plotted as the abscissa, and the precision is plotted as the ordinate to form an RP curve, which can reflect the classifier's covering ability and recognition accuracy of positive samples. The area enclosed by the curve is the Average Precision (AP) of the current category.

For the multi-classification problem, we averaged the AP for all categories, which is called mAP; it is expressed as follows:

$$\mathrm{AP} = \int_0^1 p(r)dr \tag{11}$$

$$\mathrm{mAP} = \frac{\sum\limits_{i=1}^{c} \mathrm{AP}(\mathrm{C}_i)}{\mathrm{C}} \tag{12}$$

In the formula, $p(r)$ is RP curve, C_i is category i, and C is the total category of equipment.

In addition to the mAP parameters, the test speed of the model is evaluated by Frames Per Second (FPS) and the time, which are reciprocal to each other. FPS and time are also used in this article to evaluate the capability of the model due to the rapid and accurate requirements of the engine room monitoring system.

4.2.2. Model Validation on the PASCAL VOC Dataset

In this section, in order to verify the validity of the model, this article first conducted experiments on the common dataset PASCAL VOC. Not only can the validity of the model be verified by training PASCAL VOC, but also the training weight can be obtained. This weight can be used in the training of the EMER dataset, and convergence can be achieved more quickly through transfer learning fine-tuning.

In this research, the model is trained using the training set and verification set from PASCAL VOC2007++2012, and it is tested using the test set from PASCAL VOC2007. In terms of training details, we train for 100 epochs on CBS-YOLOv5m. SGD is used as the optimizer, the learning rate is set at 0.01, the momentum is set at 0.937, the weight attenuation is set at 0.0005, and the image batch is 32.

In this paper, we compare two typical indicators: FPS and mAP. As shown in Table 3, the CBS-YOLOv5m proposed in this article is improved. Compared with current mainstream models, including Faster R-CNN [8], R-FCN [33], YOLO3 [14], YOLOv4 [2], YOLOX [34], YOLOv7 [15], SSD [9], Deconvolutional Single-Shot Detector (DSSD) [35], Attentive Single-Shot Multibox Detector (ASSD) [36], RetinaNet [10] and YOLOv8m [37], our model is improved.

Table 3. Compared with other methods on PASCAL VOC dataset.

Model	Input Size	Backbone	Train	Test	FPS	mAP (%)
Faster R-CNN	~600 × 1000	ResNet-101	VOC07+12	VOC07	2.4	76.4
R-FCN	~1000 × 800	ResNet-101	VOC07+12	VOC07	5.9	80.5
YOLOv3	352 × 352	Darknet53	VOC07+12	VOC07	19.9	75.7
YOLOv4	416 × 416	CSPDarknet53	VOC07+12	VOC07	48.1	85.7
YOLOXm	640 × 640	Darknet53	VOC07+12	VOC07	81.3	88.2
YOLOv7	640 × 640	ELANnet	VOC07+12	VOC07	98.02	88.4
YOLOv8m	640 × 640	CSPDarknet53	VOC07+12	VOC07	100.27	87.9
SSD	512 × 512	VGG16	VOC07+12	VOC07	25.2	79.8
DSSD	513 × 513	ResNet-101	VOC07+12	VOC07	5.5	81.5
ASSD	512 × 512	VGG16	VOC07+12	VOC07	35.9	81.6
RetinaNet	600 × 600	ResNet-50	VOC07+12	VOC07	17.4	79.3
YOLOv5m	640 × 640	CSPDarknet53	VOC07+12	VOC07	87.08	87.0
CBS-YOLOv5m	640 × 640	CSPDarknet53	VOC07+12	VOC07	56.05	88.8

4.2.3. Ablation Study on EMER Dataset

In this paper, the weight of training on the PASCAL VOC2007++2012 dataset is used as the pre-training weight. As for the training details, 80 epochs are trained with the Adam optimizer. The learning rate is set to 0.001, the moving average attenuation rate is 0.9, and the weight attenuation is set to 0.0005. The batch of images is 32. When using the default anchor box to train the model in this article, the mAP is 83.42%. The model is trained using the anchor boxes obtained by clustering, and the obtained mAP is 83.73. Since the subsequent experiments in this paper are based on the new anchor box, the model trained by the new anchor box in this paper is used as the baseline.

This article conducts ablation experiments on the EMER dataset to respectively explore the influence of improved backbone, improved PANet, improved head, and Soft-NMS on detection accuracy and detection speed. The backbone network's feature-extraction capabilities are improved using C3CA; VH-BiFPN is used to replace the original PANET network; and the Swin transformer is incorporated into the header network to enhance the capability of extracting global information. Finally, the Soft-NMS processing mechanism is used for processing. In Table 4, the experimental results are displayed. From the experimental results of M3 and M4, it can be known that VH-BiFPN is 0.22% higher than BiFPN. The accuracy of M2, M4, M5, M6, M7, and M8 is improved by 0.98%, 0.83%, 1.27%, 1.59%, 2.4% and 3.34%.

Table 4. EMER dataset ablation experiment. Do an experiment with the added modules here, and name them according to M1-8, where VH-BiFPN represents the weighted bidirectional feature pyramid structure of vertical and horizontal connections. SPH represents the detection head based on Swin Transformer.

Model	M	C3CA	BiFPN	VH-BiFPN	SPH	Soft-NMS	Time (ms)	mAP (%)
Baseline	1	-	-	-	-	-	20.39	83.73
Schemes	2	√	-	-	-	-	22.90	84.71 (+0.98)
	3	-	√	-	-	-	20.79	84.34 (+0.61)
	4	-	-	√	-	-	21.03	84.56 (+0.83)
	5	-	-	-	√	-	22.92	85 (+1.27)
	6	√	-	√	-	-	23.54	85.32 (+1.59)
	7	√	-	√	√	-	25.62	86.13 (+2.4)
	8	√	-	√	√	√	29.91	87.07 (+3.34)

It can be known from the experimental results that although the accuracy has increased by 3.34%, the inference time has increased by 9 ms. The increase in time is due to the introduction of the attention mechanism and Soft-NMS. This is due to the use of the self-attention mechanism, which not only requires matrix calculations but also calculates the correlation between pixels in a fixed window, resulting in an increase in the number of calculations. Although we only increase the self-attention mechanism on the head, which has a small feature map, it still brings about a 2 ms time increase. For the lightweight attention mechanism CA, the inference time is also increased by 2 ms due to the addition of convolution calculations. Compared with NMS, Soft-NMS increases the calculation of the decay score, so the post-processing process increases by 4 ms.

Figure 12 shows a comparison of the PR curves of CBS-YOLOv5m and YOLOv5m. The area contained by the curve is mAP, and the CBS-YOLOv5m area is 0.0334 greater than that of YOLOv5m.

Figure 13 is a diagram of the confusion matrix results of CBS-YOLOv5m on the EREM dataset. The results in Figure 13 show that many valves and meters are predicted as the background, with high FP and FN levels. This is because there are many small targets of valves and meters, which are difficult to be detected in dense occlusion environments. There are higher FNs for reservoir, engine, and cooler, which is due to the fact that their training samples are smaller than other types and feature extraction is limited.

4.2.4. Compared with Mainstream Models

In Table 5, the proposed model achieves a higher mAP compared to other models, which are 10.94%, 10.86%, 8.01%, 4.78%, 1.26%, 1.37%, 1.94%, 1.35% better than Faster R-CNN, SSD, ASSD, RepVGG-RetinaNet, YOLOv7, YOLOv8m, YOLOv5l, YOLOv5x and 3.34% higher than the benchmark YOLOv5m. CBS-YOLOv5m has more advantages than other models in the detection of small objects.

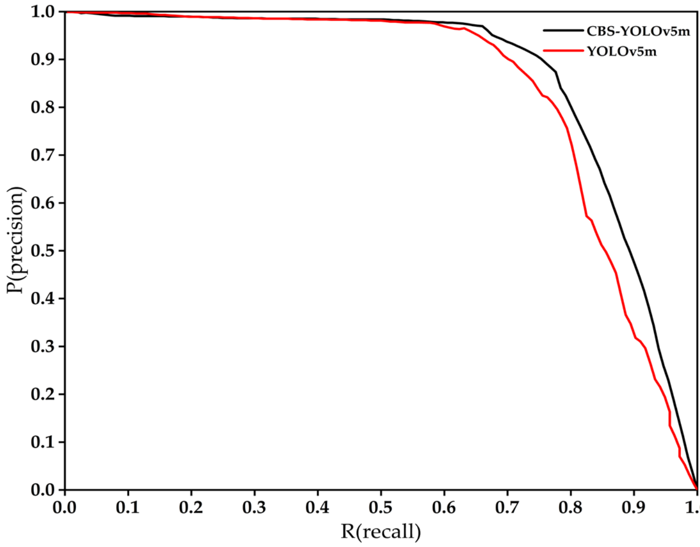


Figure 12. PR curve comparison.

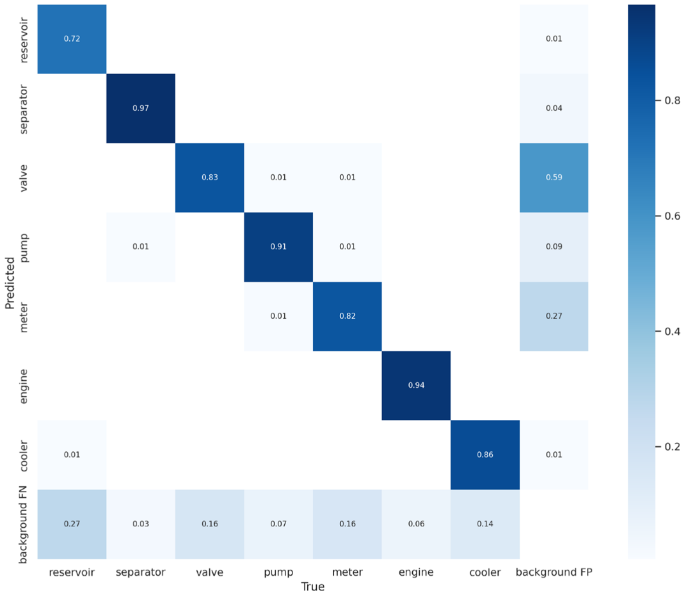


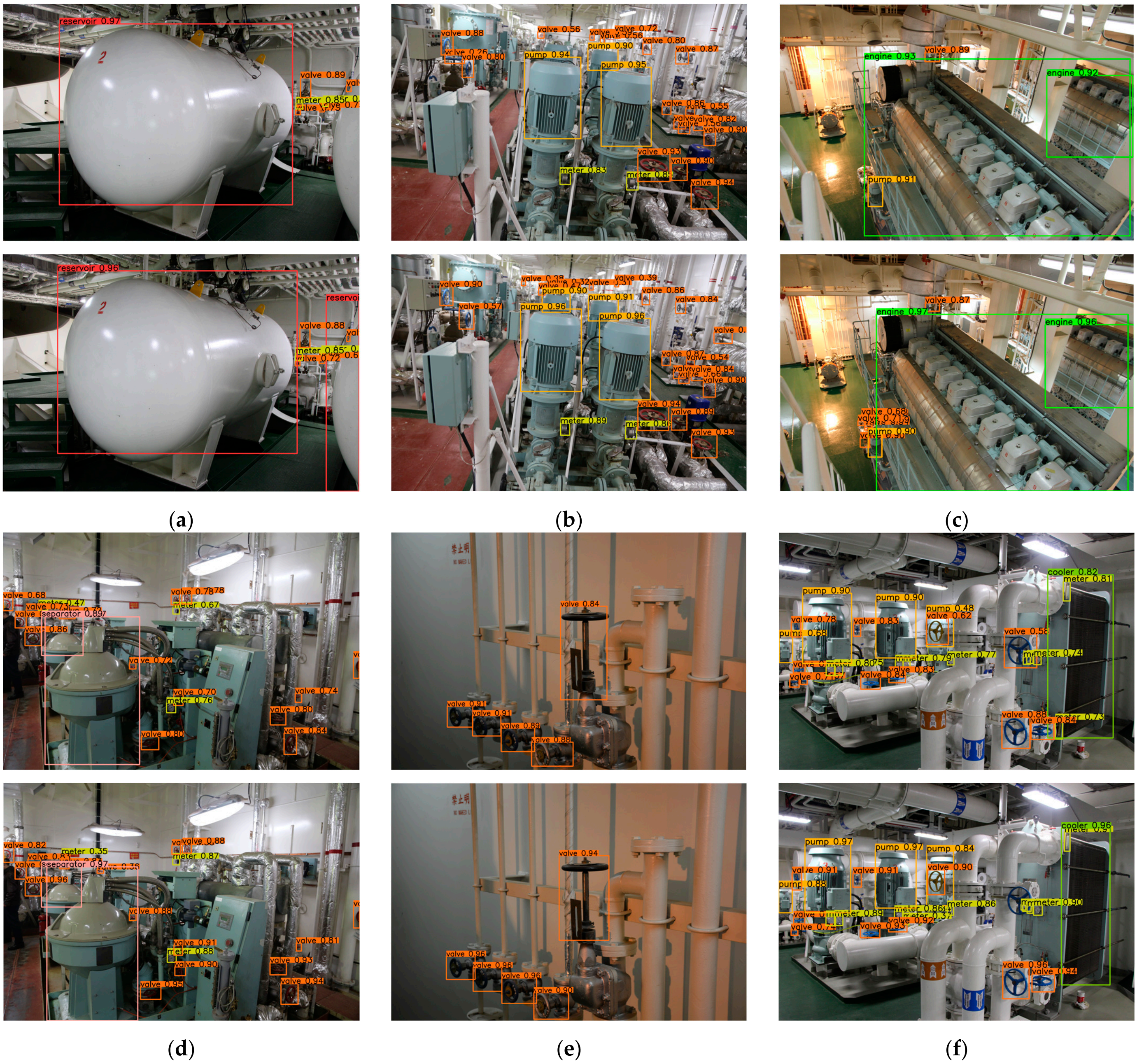
Figure 13. Confusion matrix results.

Table 5. Comparing results with other mainstream models on the EMER dataset.

Model	AP (%)							FPS	mAP (%)
	Engine	Pump	Cooler	Separator	Meter	Reservoir	Valve		
Faster R-CNN	93.77	82.11	90.96	84.83	43.81	86.95	50.49	8.53	76.13
SSD	100.00	89.46	83.53	91.71	46.22	71.05	51.48	27.99	76.21
ASSD	100.00	90.39	85.85	93.90	49.53	78.57	55.18	17.94	79.06
RepVGG-RetinaNet	100.00	95.30	93.44	97.68	60.26	55.01	74.37	24.98	82.29
YOLOv7	93.50	89.10	88.30	96.20	80.20	74.40	79.00	54.49	85.81
YOLOv8m	93.70	93.10	91.00	95.00	75.00	73.70	78.40	56.46	85.70
YOLOv5l	95.40	88.90	88.80	93.30	77.30	75.60	76.60	39.73	85.13
YOLOv5x	95.90	89.40	88.40	95.10	78.40	75.60	77.30	34.01	85.73
YOLOV5m	94.40	87.50	89.80	94.60	73.40	70.60	75.80	49.04	83.73
CBS-YOLOv5m	96.90	90.40	89.20	96.10	81.20	75.60	80.10	33.43	87.07

4.2.5. Visualization

This paper selects several typical scenes for visual comparison. As shown in Figure 14, the YOLOv5m model is compared with the detection results of the CBS-YOLOv5m model proposed in this paper. In Figure 14a, we show reservoirs. In other figures, we show the pump (b), meter (b), engine (c), separator (d), valve (e), and cooler (f). Among the six comparisons, above is the output of the original YOLOv5m, and below is the output of CBS-YOLOv5m.



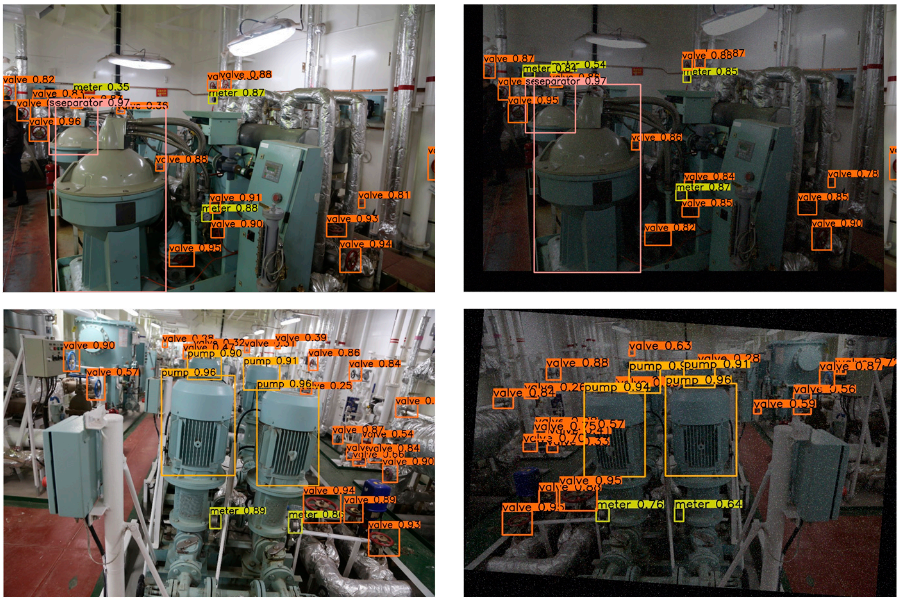
(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 14. Test set selection picture-detection results comparison.

In this paper, the IOU threshold and confidence threshold are set to 0.5 and 0.25, respectively. In Figure 14a, there are reservoirs, valves, and other equipment. CBS-YOLOv5m detects one more reservoir than YOLOv5m. In Figure 14b, CBS-YOLOv5m detects one more pump and one more valve than YOLOv5m. In Figure 14c, CBS-YOLOv5m detects several more valves than YOLOv5m with higher confidence but falsely detects an object as a valve. In Figure 14d, CBS-YOLOv5m detects one more valve than YOLOv5m, and the remaining devices have higher confidence. In Figure 14e,f, the device detected by CBS-YOLOv5m has higher confidence, but there is a missed valve in Figure 14f. From the results, the improved model in this paper can alleviate the problems of missed detection and false detection and can exceed YOLOv5m in the detection of each type of equipment. However, there are also shortcomings, such as predicting the valve as the background. In summary, CBS-YOLOv5m has a stronger recognition ability.

In order to verify the generalization ability of the model, we compared the detection effects of the same device in different lighting environments. As shown in Figure 15, column a is the detection effect of standard cabin lighting, and column a * is the detection effect of the cabin after darkening. From the perspective of the confidence of the detection effect, there is no significant change, but some small targets are missed due to the darkening of the environment. Overall, the model in this paper has strong generalization.



(**a**) (**a ***)

Figure 15. Comparison of detection results in different engine room environments. (**a**) Test results in normal environment. (**a ***) Test results in darkened environment.

5. Conclusions

To develop engine room visual perception technology, we built the EMER dataset. In order for the ship engine room robot to better identify the engine room equipment and perform related operations, this paper improves the recognition accuracy and generalization ability of the ship equipment-detection model, which enables the robot to accurately identify equipment in different environments. It also proposes an improved YOLOv5m engine room target-detection algorithm. According to the experimental results, it can be seen that: (1) The new anchor box obtained by the K-means algorithm based on the genetic algorithm can effectively improve the ability of the model to recognize the target shape and improve the detection accuracy by 0.31%. (2) The improved C3CA module can more effectively extract the main features in the image, thereby improving the detection ability of all devices. (3) The designed VH-BiFPN feature-fusion network can enhance the fusion of semantic information at different resolutions, and the mAP is increased by 0.83%. (4) Changing the convolution prediction head to the SPH detection head can calculate the global correlation, thereby avoiding the loss of features of devices with small image proportions, and can improve the problems of false detection and missed detection. (5) The Soft-NMS post-processing mechanism can effectively solve the problem of missed detection In dense scenes. The improved CBS-YOLOv5m in this paper is 3.34% higher than the mAP of YOLOv5m. However, the inference time increased by 9 ms, and the FPS decreased by 16.

The purpose of this article is for the results to be deployed in engine room monitoring and with engine room inspection robots. The main task deployed to the robot is to accurately find the location of the equipment in the complex engine room environment. Therefore, this paper focuses on the improvement of detection accuracy.

The improved model in this paper can realize the identification of cabin equipment but has the following shortcomings: (1) The current dataset only has images of equipment appearance, and images containing equipment defects need to be collected. (2) There is a problem of data imbalance among different equipment types, which leads to the low single-class accuracy of reservoirs. (3) The number of meters and valves in the engine room is large. They are densely distributed, and the targets are small, which leads to meters and valves having a high FPs. (4) The improved model has the highest mAP, but various devices' detection effects and accuracy are not the best. Therefore, multi-model integration can be considered for device detection. (5) The proposed model only considers the accuracy improvement, and the speed is lost by nearly 35%. Subsequent work should balance speed and accuracy. (6) The dataset in this paper uses static images, and it is yet to be possible to test the effects of vibration and light changes. Relevant videos should be collected in future work to better approach the real ship environment. In the future, zero-sample detectors should be studied so that the model can learn autonomously and be deployed in cabin robots to empower intelligent engine rooms.

Author Contributions: Conceptualization, Y.W. and J.Z. (Jundong Zhang); methodology, Y.W.; software and experiments, Y.W.; validation, Y.W. and J.Z. (Jinting Zhu); formal analysis, Y.W.; investigation, Y.W. and G.Z.; resources, J.Z. (Jundong Zhang;) data curation, Y.W.; writing—original draft preparation, Y.W., Y.G. and J.Z. (Jinting Zhu); writing—review and editing, J.Z. (Jundong Zhang); visualization, Y.W. and J.Z. (Jinting Zhu); supervision, J.Z. (Jundong Zhang); project administration, J.Z. (Jundong Zhang); funding acquisition, J.Z. (Jundong Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by High-technology Ship Research Program (CBG3N21-3-3).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Processed data cannot be shared at this time as they are also part of ongoing research.

Acknowledgments: The original images of the dataset EMER are provided by the 3D Virtual Engine Room team, thanks to their support for this research.

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

References

1. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
2. Bochkovskiy, J.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
3. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
5. 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.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
7. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE. Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
9. 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.

10. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
11. Lin, T.Y.; Dollar, 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. 936–944.
12. 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, 27–30 June 2016; pp. 779–788.
13. 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. 6517–6525.
14. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
15. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
16. Qi, J.; Zhang, J.; Meng, Q. Auxiliary Equipment Detection in Marine Engine Rooms Based on Deep Learning Model. *J. Mar. Sci. Eng.* **2021**, *9*, 1006. [CrossRef]
17. Shang, D.; Zhang, J.; Zhou, K.; Wang, T.; Qi, J. Research on the Application of Visual Recognition in the Engine Room of Intelligent Ships. *Sensors* **2022**, *22*, 7261. [CrossRef] [PubMed]
18. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Adv. Neural Inf. Process Syst.* **2015**, *28*, 2017–2025.
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
20. 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.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
22. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 558–567.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
24. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
25. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. *arXiv* **2021**, arXiv:2111.12419.
26. He, K.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. 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.
28. Tan, M.X.; Pang, R.M.; 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.
29. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2778–2788.
30. Hua, F. Improved Surface Defect Detection of YOLOV5 Aluminum Profiles based on CBAM and BiFPN. *Int. Core J. Eng.* **2022**, *8*, 264–274.
31. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition, HongKong, China, 20–24 August 2006; pp. 850–855.
32. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS - Improving Object Detection With One Line of Code. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5562–5570.
33. Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. *Adv. Neural Inf. Process Syst.* **2016**, *29*, 379–387.
34. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
35. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
36. Yi, J.; Wu, P.; Metaxas, D.N. ASSD: Attentive Single Shot Multibox Detector. *Comput. Vis. Image. Underst.* **2019**, *189*, 102827. [CrossRef]
37. Reis, D.; Kupec, J.; Hong, J.; Daoudi, A. Real-Time Flying Object Detection with YOLOv8. *arXiv* **2023**, arXiv:2305.09972.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.