electronics

MDPI

Article

An Underwater Dense Small Object Detection Model Based on YOLOv5-CFDSDSE

Jingyang Wang [1,2], Yujia Li [1], Junkai Wang [1] and Ying Li [1,*]

[1] Hebei University of Science and Technology, Shijiazhuang 050018, China
[2] Hebei Intelligent Internet of Things Technology Innovation Center, Shijiazhuang 050018, China
* Correspondence: jszxly@hebust.edu.cn

Abstract: Underwater target detection is a key technology in the process of exploring and developing the ocean. Because underwater targets are often very dense, mutually occluded, and affected by light, the detection objects are often unclear, and so, underwater target detection technology faces unique challenges. In order to improve the performance of underwater target detection, this paper proposed a new target detection model YOLOv5-FCDSDSE based on YOLOv5s. In this model, the CFnet (efficient fusion of C3 and FasterNet structure) structure was used to optimize the network structure of the YOLOv5, which improved the model's accuracy while reducing the number of parameters. Then, Dyhead technology was adopted to achieve better scale perception, space perception, and task perception. In addition, the small object detection (SD) layer was added to combine feature information from different scales effectively, retain more detailed information, and improve the detection ability of small objects. Finally, the attention mechanism squeeze and excitation (SE) was introduced to enhance the feature extraction ability of the model. This paper used the self-made underwater small object dataset URPC_UODD for comparison and ablation experiments. The experimental results showed that the accuracy of the model proposed in this paper was better than the original YOLOv5s and other baseline models in the underwater dense small object detection task, and the number of parameters was also reduced compared to YOLOv5s. Therefore, YOLOv5-FCDSDSE was an innovative solution for underwater target detection tasks.

Keywords: underwater object detection; YOLOv5; CFnet; SD; SE

Citation: Wang, J.; Li, Y.; Wang, J.; Li, Y. An Underwater Dense Small Object Detection Model Based on YOLOv5-CFDSDSE. *Electronics* **2023**, *12*, 3231. https://doi.org/10.3390/electronics12153231

Academic Editors: Tomasz Trzcinski and Beiwen Li

Received: 20 June 2023
Revised: 13 July 2023
Accepted: 24 July 2023
Published: 26 July 2023

1. Introduction

The ocean is the largest ecosystem on the Earth, covering approximately 71% of the Earth's surface area. With the continuous improvement of scientific and technological productivity, human beings are constantly promoting the development and utilization of marine resources, and underwater target detection technology comes into being in this process. Nowadays, underwater target detection technology has become an important research direction in computer vision and artificial intelligence, and it has extensive value in many practical applications, such as marine ecological protection, underwater archaeology, seabed resource exploration, and underwater pipeline detection [1]. Applying this technology well can not only save a lot of human resources and material resources but also bring people into the vast and mysterious ocean world.

In recent years, researchers used traditional feature extraction and pattern recognition algorithms, such as edge detection, Hough transform, and template matching, for target detection [2–4]. However, there are some limitations and shortcomings in the traditional image processing methods, which lead to the inability to capture the complex information in the image fully, cannot adapt to these complex scenes well, and require more computing resources and time. With the rise of machine learning, researchers began to use machine learning methods for target detection, such as support vector machines, random forests, and K-nearest neighbor [5–7]. Although machine learning methods made some achievements

in target detection tasks, the generalization ability of machine learning models is limited, and specific adjustments are required in each particular task. This adjustment process often requires a lot of domain knowledge and experience. With the vigorous development of deep learning technology, many defects of traditional image processing and machine learning methods in target detection have been successfully solved. Currently, deep-learning-based target detection techniques are mainly divided into two categories. The first category is typified by the R-CNN family of algorithms [8], known as two-stage methods, and its core idea is based on candidate regions. The second category is a one-stage method, mainly including RetinaNet [9] and YOLO [10] series algorithms, among which YOLO developed rapidly in recent years. The two-stage method primarily performs classification and regression on a series of sparse candidate boxes using techniques like selective search, which makes the model achieve higher accuracy. The one-stage approach has a high computational efficiency along with a fast classification speed due to its ability to densely sample different locations of the image at different scales and proportions. The one-stage method also uses convolutional neural networks (CNN) to extract features from the image and eventually classify the objects.

Underwater target detection faces many problems that do not exist on land. First of all, the characteristics of light scattering and absorption in the underwater optical environment often lead to the degradation of the quality of the collected images, the loss of details, and a reduction in contrast. Secondly, because some creatures have mimic characteristics and some creatures have the habit of living in groups, the creatures in the collected underwater images often have problems of occlusion and denseness. At the same time, many underwater organisms are small objects, which makes it more difficult to detect the collected images. Due to these problems, the existing underwater object detection algorithms need to be optimized to obtain better detection results. The definition of small objects in the paper is usually determined according to the specific data set. Zhu et al. [11] defined objects whose width accounts for less than 20% of the entire image as small objects in their dataset. In the COCO dataset, pixels smaller than 32×32 are called small objects; Liu et al. [12] counted objects with an area of less than 5% in their underwater target data set called small objects. In this paper, the detection object area accounts for less than 5% of the entire image area as a small object. This paper proposes a new underwater target detection model YOLOv5-CFDSDSE for the problems existing in the underwater environment, which is improved based on YOLOv5s. The YOLOv5-CFDSDSE adopts a new CFnet method, introduces the Dyhead method, adds an SD layer, and uses the SE attention mechanism. These improved methods optimize the backbone, neck, and head of YOLOv5s.

The contributions of this paper are as follows:

(1) A new structure Cfnet is proposed, which is an efficient fusion of C3 and FasterNet structures, reduces the number of parameters, increases the detection speed, and, at the same time, has a high detection accuracy;
(2) The Dyhead replaces the original detection head, improving the model's ability to detect multi-scale and multi-category targets, which is especially effective for underwater dense small object detection;
(3) The SD layer is added, improving the model's performance for underwater small object detection;
(4) The attention mechanism SE is introduced to make the model capture the global context information more completely.

2. Related Work

Recently, there were many research achievements in underwater target detection. Chen et al. [13] conducted a study in underwater object detection, using a combination of visual features such as color and light transmission information to explore its potential in this field. Initial recognition regions were generated using visual features and optical transmission information, and these regions were further optimized by using image segmentation techniques to finally obtain the detection results for underwater targets. How-

ever, the detection results are unstable due to the extreme complexity of the underwater optical environment. Chen et al. [14] proposed a multiscale Retinex enhancement algorithm that combined the Retinex algorithm by emulating the fish retina to eliminate underwater noise. They also used deep learning methods to improve the detection performance of small objects. However, the Retinex algorithm is prone to the halo phenomenon in the transition region of strong light and shadow, which weakens image details. Liu et al. [15] proposed an underwater target-detection algorithm based on Faster RCNN. The algorithm used Swin Transformer to replace the backbone network of Faster RCNN and introduced a path aggregation network to achieve the fusion of multi-level feature maps. The ROI pooling was also improved to ROI align, thus improving the detection performance of the algorithm. However, the algorithm's network parameters are large, making its deployment difficult.

Wei et al. [16] proposed an improved YOLOv3 model, which aimed to enhance the semantic information of depth features and improve the performance of small object detection. The model introduced an attention mechanism after the deep convolutional layer and combined deep semantic information with shallow position information. In this way, the model can capture the semantic details of the target more accurately and improve the detection of small objects. However, this increased the computational complexity of the model, increasing training and inference time, thus posing certain challenges for real-time applications or resource-constrained scenarios. Muksit et al. [17] proposed a specific YOLO-Fish algorithm for underwater fish target detection. The researchers proposed two models in this algorithm, namely YOLO-Fish-1 and YOLO-Fish-2. YOLO-Fish-1 reduced the false detection of tiny fish by improving the upsampling step size. YOLO-Fish-2 enhanced the ability to detect fish dynamically by introducing a spatial pyramid structure. This study's innovation lay primarily in optimizing tiny fish and dynamic scenes. However, implementing the algorithm required more computing resources and affects operating efficiency. Zhang et al. [18] proposed an improved underwater target-detection model based on the YOLOv4. The model used a K-means++ clustering method to optimize the anchor frame, introduced an additional detection head to handle targets of different sizes, and, finally, used the FIoU loss function to replace the traditional loss function. However, the new approach led to more extended training and inference time for the algorithm.

Shi et al. [19] proposed an improved YOLOv5 algorithm. The optical and thermal images were enhanced by data enhancement technology, which improved the generalization ability of the detection algorithm. A linear feature detection technique was introduced, which enhanced feature propagation and improved feature utilization. However, in a complex underwater environment, the detection accuracy will be reduced due to the interference of factors such as water flow, clutter, and target speed. Li et al. [20] made improvements based on the YOLOv5. The CA attention mechanism and C3 module were fused into the C3CA module, which replaced the C3 module in the benchmark model to improve target feature information's extraction and detection accuracy. Using the EIOU loss function instead of the GIOU loss function optimized the algorithm's the localization accuracy and convergence speed. However, in the underwater environment, the posture of the fish may change, and the fish in the group may be overshadowed by each other, and so, the accuracy of the algorithm in this respect will be affected. Li et al. [21] used the ShuffleNetv2 lightweight network to replace the CSPDarkNet53 backbone network of YOLOv5, reducing the model's size and calculation and improving the detection speed. The PANet network was replaced with the improved BiFPN-Short network, and the improved network was used for feature fusion, which enhanced the information dissemination between different levels, thereby improving the accuracy of the detection algorithm. However, ShuffleNetv2 took more time to complete forward propagation in the inference phase, and so, its inference delay was relatively high. Wang et al. [22] proposed a YOLOX-based underwater target detection algorithm B-YOLOX-S. The wavelet transform was used to transform the style of the image, which improved the clarity of the image and the detection target, and enhanced the model's generalization ability. The method combined FPN with BIFPN-S to fuse the features of the backbone layer and accelerate model detection. The EIOU loss function

was used to improve the localization accuracy of detection. However, since the EIOU loss function needed to calculate the intersection ratio between the prediction frame and the real frame as well as the area of the overlap region between the two frames, each prediction frame needed to perform these calculations, and so, the computational complexity can be quite high in a large-scale target detection task, leading to an increase in training time.

3. Methodology

3.1. YOLOv5

The BottleNeck structure in YOLOv5 [23] is an important component used to build the model, which is a residual block aiming to improve the model's representation ability and performance. The BottleNeck can improve the representation ability of the model while maintaining a small computational burden and parameter quantity. It can capture richer and more complex features, which can help improve the performance of tasks such as object detection and image classification. The CBS consists of convolution, batch normalization, and the SiLU activation function. The local features of the image are extracted by the convolution layer, and then normalized by batch normalization. Finally, nonlinear features are introduced through the SiLU activation function, so the model can learn more complex feature representations. The combination of this structure helps improve the model's performance and generalization ability, and provides more accurate prediction results in object detection tasks. The C3 structure is an innovative design proposed by the YOLOv5 team, which aims to improve the effect of feature extraction and model performance. This cross-stage design allows feature maps of different scales to interact with each other, thus improving the model's receptive field and information transfer ability. SPPF is a technique that utilizes spatial pyramid pooling and flexible aggregation to extract feature representations with multi-scale information to improve the detection ability of object-detection models for objects of different sizes. The structure diagram of YOLOv5 model is shown in Figure 1.

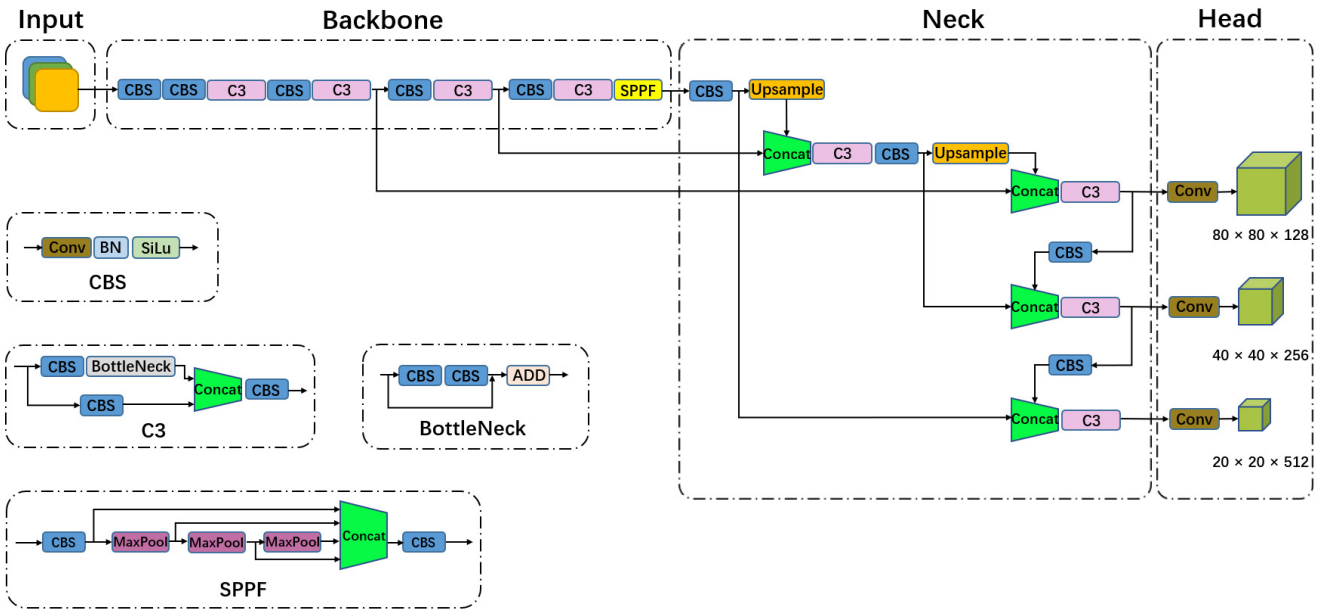


Figure 1. YOLOv5 structure diagram.

3.2. CFnet

The C3 structure plays an essential role in YOLOv5. However, since the C3 structure contains multiple convolutional layers, many convolution calculations are required, resulting in a large amount of calculation for the model. The inference speed of the model is affected, especially in resource-constrained environments. FasterNet [24] is an efficient neural network designed for target detection and localization tasks, which can solve the

computational insufficiency of the C3 structure and reduce the computational load of the model. Its advantages are particularly obvious in resource-constrained environments. Partial Convolution (PConv) is a competing alternative to reduce computational redundancy and the number of memory accesses, exploit the redundancy in feature maps, and systematically apply Conv on only a part of the input channels without affecting the remaining channels. FasterNet consists of four hierarchical stages. More FasterNet blocks are placed, and more computational tasks are assigned in the last two stages since they consume less memory access. Each FasterNet block has a PConv layer followed by a Conv 1×1 layer. Normalization and activation layers are indispensable for high-performance neural networks. However, overuse of these layers can limit the diversity of features, affecting performance and eventually leading to slower calculation speed. FasterNet uses these layers only after partial convolutions, achieving lower latency and preserving feature diversity. In terms of normalization layers, the network uses the Batch Normalization (BN) method to achieve faster inference speed. For the activation layer, FasterNet uses GELU for the smaller FasterNet variant and ReLU for the larger FasterNet variant. The last three layers of the network structure, namely global average pooling, 1×1 Conv and fully connected layers, are used for feature transformation and classification. The structural diagram of FasterNet is shown in Figure 2. The contents of the dot box represent a module, and gray dotted lines represent the structures' explanations.

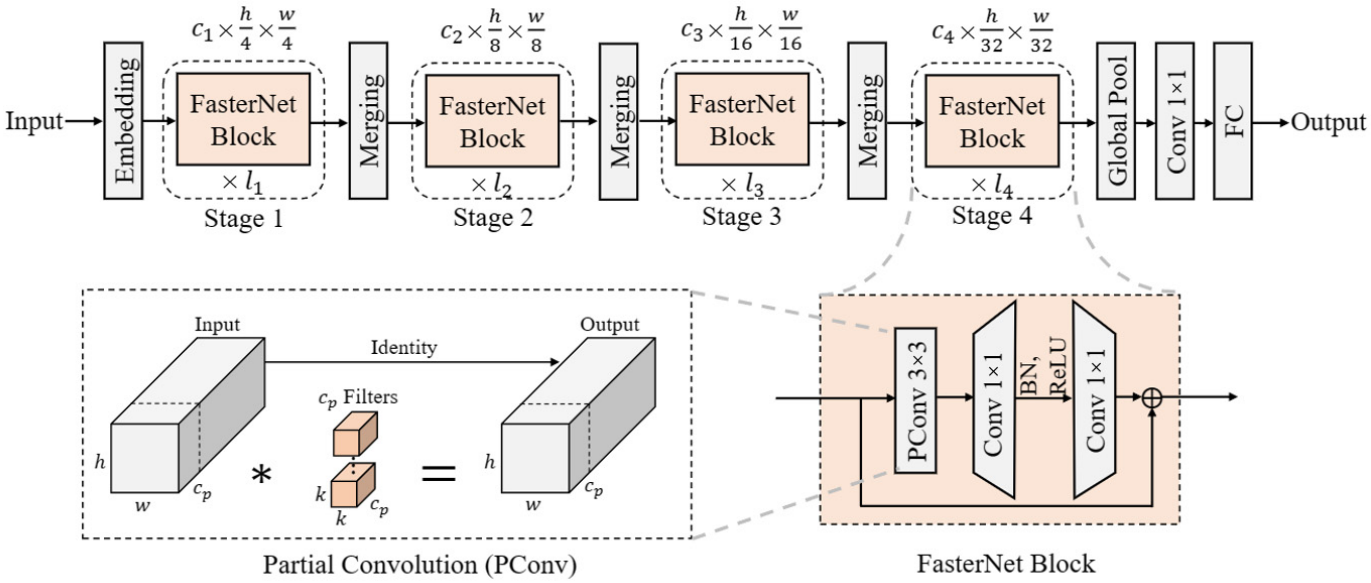


Figure 2. Schematic diagram of FasterNet structure.

In this paper, the CFnet structure was proposed using the idea of FasterNet, and the results of C3 output were used as input. Firstly, the C3 output was shortcut connected. This connection mechanism can directly transfer information from earlier layers to subsequent layers, which is helpful for information transfer and gradient flow. Next, the result of the shortcut connection was input into the PConv operation. Then, MLP operation was carried out on the results after PConv operation. In this paper, the MLP consisted of a CBS and a Conv2D operation. These operations further processed the features, enhancing their expressiveness and perception. Finally, the CFnet structure concatenated the result after the MLP operation with the content of the previous shortcut to obtain the final output. This connection operation can comprehensively utilize early and late feature information to improve the expressive ability and detection performance of the model.

The CFnet structure extracted and enhanced features from the output of C3 by applying shortcut connections, PConv operations, MLP and other operations to obtain more accurate and comprehensive object detection results. This design considered the combined

use of feature transfer and enhancement, MLP application, and connection operations to provide an innovative solution for object detection tasks. CFnet is shown in Figure 3.

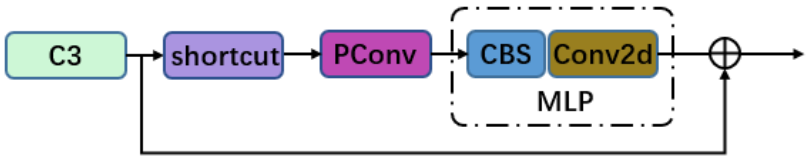


Figure 3. Cfnet structure diagram.

3.3. Dyhead

The detection head of YOLOv5 is a multilayer convolutional and fully connected network responsible for extracting features from images and performing object detection and localization. However, YOLOv5 cannot detect and localize small objects effectively because of the larger field of perception and lower resolution, and so, it has some difficulties in detecting small objects.

Dyhead [25] is a novel dynamic head framework that aims to improve the performance of localization and classification in object detection tasks. Dyhead adopts an attentional mechanism perspective, combining the target detection head with an attentional mechanism. By cooperatively combining multiple self-attention mechanisms, scale awareness, space awareness, and task awareness are achieved between feature layers, spatial locations, and within task channels. This combined attention mechanism significantly improves the representation of the target detection head with no additional computational overhead. Experiments on the COCO dataset show that Dyhead achieved impressive performance in the object-detection task, and its detection's Average Precision (AP) reached 60.6. The structural diagram of Dyhead is shown in Figure 4.

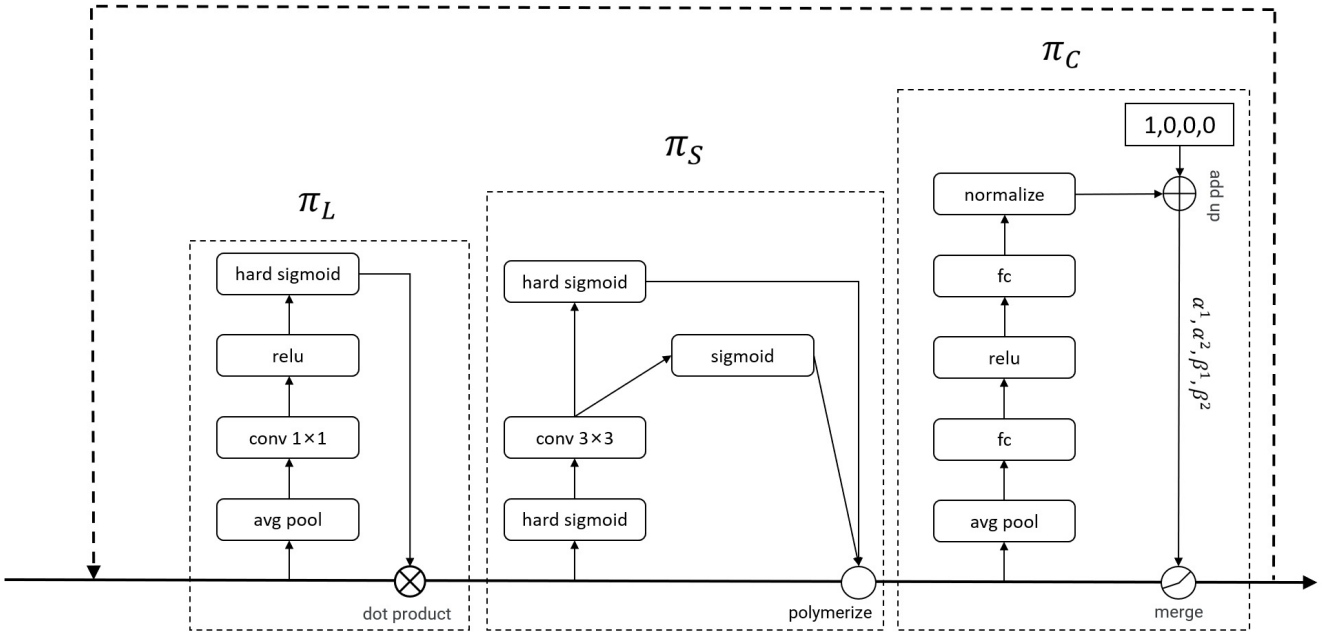


Figure 4. Dyhead structure diagram.

In Figure 4, π_L is scale-aware attention, and it is formulated as:

$$\pi_L(\mathcal{F}) \cdot \mathcal{F} = \sigma\left(f\left(\frac{1}{SC}\sum_{S,C}\mathcal{F}\right)\right) \cdot \mathcal{F}, \quad (1)$$

In the formula, the feature pyramid can be expressed as a 4-dimensional tensor $\mathcal{F} \in \mathcal{R}^{L\times H\times W\times C}$. L indicates the number of layers in the pyramid. H, W, and C denote the

height, width, and number of median-level feature channels, respectively. Furthermore, we define $S = H \times W$, reshape the tensor into a three-dimensional tensor $\mathcal{F} \in \mathcal{R}^{L \times S \times C}$, $f(\cdot)$ is a linear function approximated by a 1×1 convolutional layer. This means that the algorithm uses a 1×1 convolution operation to approximate a linear function. A 1×1 convolution considers only itself at each pixel location, and linearly transforms the input through the weights in the convolution kernel. $\sigma(x) = \max\left(0, \min\left(1, \frac{x+1}{2}\right)\right)$ is a hard-sigmoid function. The hard-sigmoid function is a nonlinear function that maps input values to a range between 0 and 1. This function is simpler and more efficient to implement, thus reducing the amount of computation and improving performance.

π_S is spatial-aware attention, and it is formulated as:

$$\pi_S(\mathcal{F}) \cdot \mathcal{F} = \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{K} w_{l,k} \cdot \mathcal{F}(l; p_k + \Delta p_k; c) \cdot \Delta m_k, \tag{2}$$

π_S is a two-step approach, which first makes the attention learning sparse by using deformable convolution [26] and then aggregates features across levels at the same spatial locations. The content in parentheses is responsible for the convolution. k represents the number of sparsely sampled locations. p_k represents a sampling position, and $p_k + \Delta p_k$ represents the position moved by the self-learned spatial offset Δp_k to focus on a discriminative region. Δm_k represents the self-learning importance scalar at position p_k. These two parameters are learned from the input features of the middle layer of $\mathcal{F}$. The model can selectively focus on discriminative regions in the input features and assign different importance weights to each location through the learned spatial offset and importance scalar. The ';' in the formula separates multiple variables to indicate that they are distinct entities.

π_C is task-aware attention, and it is formulated as:

$$\pi_C(\mathcal{F}) \cdot \mathcal{F} = \max\left(\alpha^1(\mathcal{F}) \cdot \mathcal{F}_c + \beta^1(\mathcal{F}), \alpha^2(\mathcal{F}) \cdot \mathcal{F}_c + \beta^2(\mathcal{F})\right), \tag{3}$$

$\mathcal{F}_c$ represents the feature slice on the cth channel. $\left[\alpha^1, \alpha^2, \beta^1, \beta^2\right]^T = \theta(\cdot)$ is a hyper-function for learning to control the activation threshold. The implementation of $\theta(\cdot)$ is as follows [27]: first, perform global average pooling on the $L \times S$ dimension to reduce the dimensionality, and then, use two fully connected layers and a normalization layer and, finally, apply a translated sigmoid function to normalize the output to the range $[-1, 1]$. By learning the obtained hyperparameters, the model can adaptively adjust the activation threshold to control the activation degree of the features.

3.4. SD

In the Neck part of YOLOv5, the characteristic information of small objects is easily diluted during the propagation process in the network. Therefore, this paper designed and implemented a new small object detection layer SD, which focused on capturing and retaining the feature information of these small objects, enabling the model in this paper to detect small objects more accurately.

SD was located after the second upsampling, and the feature maps after the second upsampling usually had higher resolution and richer semantic information. Placing the small object detection layer after the second upsampling can better use these high-resolution features, and obtain more global and local contextual information by fusing low-level and high-level features. At the same time, because small objects occupied a relatively small pixel level in the image, they were easily distracted by the surrounding background. Placing the small object detection layer after the second upsampling allowed the model to focus more on detecting and processing small objects without losing the detailed information of small objects on the lower-resolution feature maps. In summary, compared with the YOLOv5, the improved model had an additional feature layer with a resolution of 160×160 and a channel number of 64, which can effectively detect small objects. The comparison diagram

is shown in Figure 5. The red box represents the newly added small object detection layer, and the yellow lines represent the sampling operations.

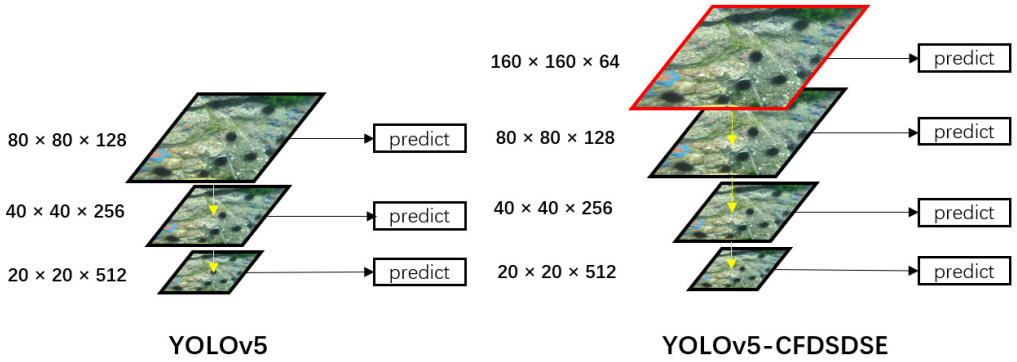


Figure 5. Comparison of model feature layers.

3.5. SE Attention

Attentional mechanism was used to dynamically allocate and focus attentional resources in neural networks to improve the model's ability to attend to and process inputs.

This paper introduced the channel attention mechanism SE [28] with the input dimension $C \times H \times W$ of the feature map, where C means the channel number of the input data, and H and W represent the height and width of the feature map, respectively. First, F_{tr} is a Transformation operation that converts the input feature map X into a feature map U. Then, the compression operation F_{sq} (Squeeze) is performed to compress the input feature map in spatial dimensions from $C \times H \times W$ to $C \times 1 \times 1$ by the global average pooling (GAP) operation. The compression features represent the global information of each channel. Then, the excitation operation F_{ex} (Excitation) is performed, where two fully connected layers (or convolutional layers) are introduced to learn the relationship between channels on the base of the compressed features. After F_{ex}, the scaling operation F_{scale} (Scale) is performed, and the importance weight of the channel is obtained, which is applied to the original feature map. The feature maps can be scaled at the channel level by multiplying the weights of each channel with the corresponding feature maps. In this way, the model can adaptively enhance the representation ability of important features. Finally, the feature map after the scaling operation is used as the final output, which can be used for subsequent tasks such as classification, target detection, etc. The structure diagram of SE is shown in Figure 6.

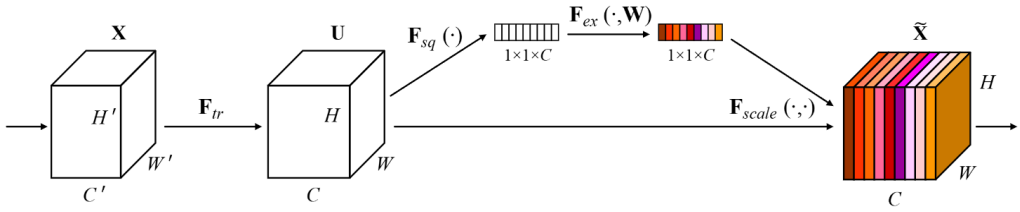


Figure 6. SE structure diagram.

The newly added SE module was located at the deepest part of the model, so the features that it operates on were the result of multi-layer convolution and feature connections. These features contain rich semantic information and high-level representations of objects. With the SE module, the model can learn which features are important and which are not, and adjust their importance accordingly. It helps the model recognize objects in complex scenes. Due to the complex underwater environment and numerous interfering information, the SE module also suppressed features that were not important, such as those that did not contribute to target detection or even interfere with detection. This helps to reduce the misclassification rate of the mode. By adaptively recalibrating the importance of

features, the SE module can improve the generalization ability of the model, so that the model can perform well in different scenarios and targets.

3.6. YOLOv5-CFDSDSE

This paper optimized the structure of the object detection model. Since C3 in YOLOv5 has a deeper network structure and more parameters than the traditional Darknet network, it requires more computing resources and longer training time during the training process. It may bring certain challenges to some scenarios with limited computing resources. In order to make the detection model have higher detection accuracy, lighter model volume and faster detection speed, a new structure CFnet was proposed. CFnet structure can avoid frequent memory access to operators in the deep convolution process, thereby reducing redundant calculations and memory access, extracting spatial features more efficiently, and improving detection accuracy. The new detection header Dyhead was used to replace the original detection header, which improved the performance of multi-scale and multi-category target detection, simplified the model structure, and improved the generalization ability. By adding a special detection layer for underwater small object detection, the features of small objects can be extracted and enhanced, thus improving detection performance, preserving high-resolution feature information, and realizing multi-scale feature fusion. The attention mechanism SE was integrated into the feature enhancement network, and the feature map's channel weight was adaptively adjusted to capture the global context information to improve the model's performance in computer vision tasks and better recognize targets. The overall structure of YOLOv5-CFDSDSE is shown in Figure 7. The red dotted box is the specific content of the SD structure.

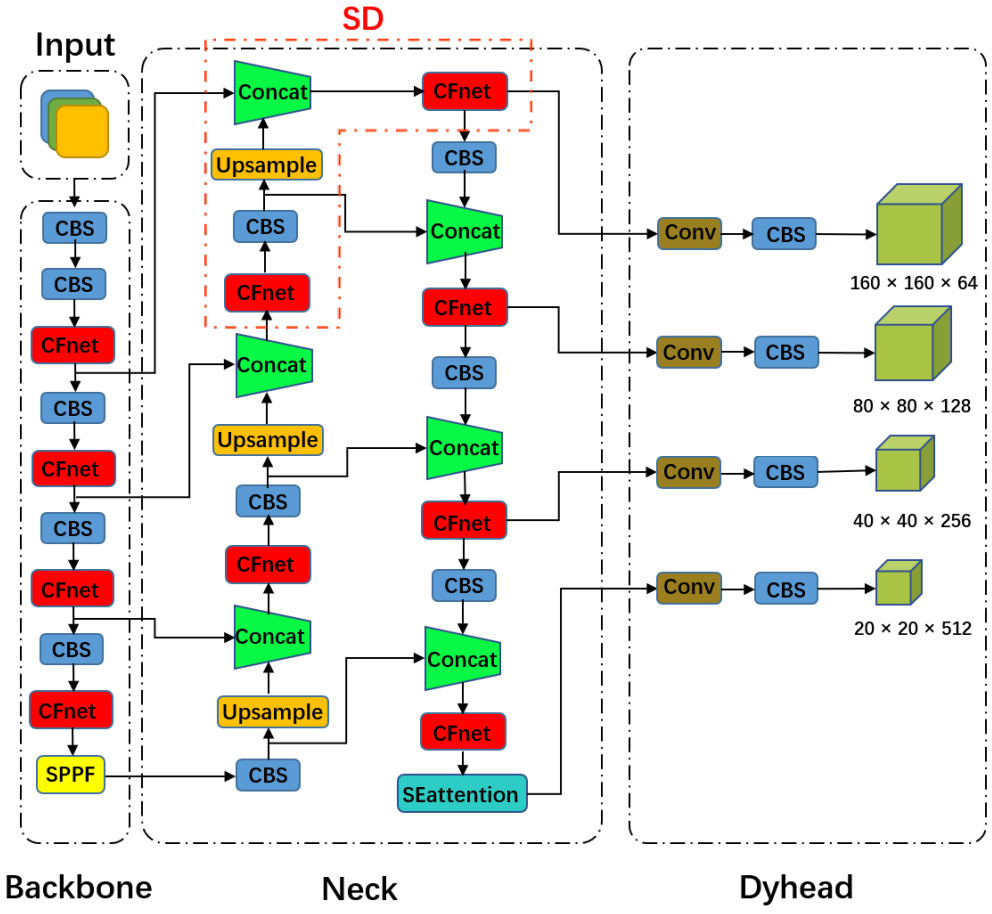


Figure 7. Overall network structure diagram.

4. Experiment

4.1. Experimental Environment

To deeply explore and validate the YOLOv5-CFDSDSE object detection model proposed in this paper, comparison and ablation experiments were conducted. All experiments were performed on the same 3080 Ti graphics card equipped with 12 G video memory and 9-core Intel(R) Xeon(R) CPU E5-2686 v4 CPU. Cuda version was 11.1.1. According to the size of the data set images, the resolution of the model input image was uniformly set to 640 × 640 pixels, and the batchsize was set to 32 for stable batch normalization and further prevention of over-fitting. The cosine annealing learning rate adjustment strategy was used, and the initial learning rate and learning rate attenuation factor were set to 0.01. All comparison experiments were trained for 400 cycles, achieving complete convergence of the experimental results.

4.2. Experimental Data Set

URPC 2019 and UODD underwater object detection datasets have classical underwater object detection scenarios, which are typical and generalizable for research. Underwater Robot Perception Challenge 2019 (URPC 2019) is a challenge for underwater robot perception, mainly aimed at promoting the development of visual perception and target recognition capabilities of underwater robots in complex environments. The Underwater Object Detection Dataset (UODD) data set [29] is a standard real-world underwater object detection dataset proposed by Dalian University of Technology. This paper aimed to research underwater dense small objects, which requires a high proportion of underwater dense objects; thus, a part of object dense image data was selected from these two datasets to form a self-made dataset (URPC_UODD), which had 3673 images and 25,122 objects. In this paper, it was randomly divided into a training set of 3000 images, a validation set of 337 images, and a test set of 336 images according to the ratio of 8:1:1. There were 16,996 objects in the training set, 4030 objects in the validation set, and 4096 objects in the test set. Most of the objects in the URPC_UODD were small, and 81% of the object boxes were less than 5% of the image area.

4.3. Comparison Experiment

In this paper, the YOLOv5-CFDSDSE was compared with some advanced underwater object detection models, including RetinaNet, Faster-RCNN, YOLOX, YOLOv7, and YOLOv5s. For YOLOX, the tiny version was chosen in this paper to make it similar in size to the YOLOv5-CFDSDSE. When using RetinaNet, this paper used EfficientNet as the infrastructure. For Faster-RCNN, this paper chose ResNet18 as the lightweight architecture. For YOLOv5, this paper chose YOLOv5s model with CSPDarknet 53 as the infrastructure. When using YOLOv7, this paper chose the lightweight YOLOv7-tiny model. This paper also added the experimental results of CenterNet and SSD to make our comparison more comprehensive. The performance metrics of these models were evaluated on the URPC_UODD dataset, and all training settings were consistent. Table 1 shows the comparison results of each model in terms of calculation consumption and model size.

Table 1. Comparison experiments of different detection models.

Model	Backbone	mAP@0.5	Precision	Recall	Number of Parameters
RetinaNet	EfficientNet	60.1%	58.0%	62.8%	37.5 M
CenterNet	ResNet18	74.1%	73.2%	75.3%	30.21 M
Faster-RCNN	ResNet18	74.7%	53.4%	83.0%	47.60 M
SSD	MobileNetV3	76.3%	75.8%	79.2%	4.92 M
YOLOX-tiny	Darknet53	78.4%	80.3%	75.8%	5.70 M
YOLOv5s	CSPDarknet53	80.8%	85.5%	75.8%	7.03 M
YOLOv7-tiny	CSPDarknet53	82.1%	82.6%	76.7%	6.02 M
YOLOV5-CFDSDSE	CSPDarknet53	85.1%	86.7%	80.2%	6.52 M

Compared with the Comparison models, it was found that the YOLOv5-CFDSDSE model proposed in this paper had slightly more parameters than YOLOX-tiny and YOLOv7-tiny. However, its accuracy increased by 6.7% and 3.0%, respectively. Compared with YOLOv5s, YOLOv5-CFDSDSE increased mAP@0.5 by 4.3% and reduced the number of parameters by about 7.8%. YOLOv5-CFDSDSE also had advantages over RetinaNet, CenterNet, SSD, and Faster-RCNN. Therefore, YOLOv5-CFDSDSE was a better choice in underwater object detection.

4.4. Ablation Experiment

This paper conducted ablation experiments to analyze the effect of each of the following improvements: CFnet module, more advanced Dyhead, SD structure, and SE attention mechanism at the deepest part of the model. Table 2 shows the results of the ablation experiments. In the table, '×' indicates that the module is not used, and '√' indicates that the module is used.

Table 2. Ablation experiments of different detection models.

Case	CFnet	Dyhead	SD	SE	mAP@0.5	mAP@0.5:0.95	Training Time	Inference Time	SODR	Number of Parameters
1	×	×	×	×	80.8%	47.2%	1.42 h	1.5 ms	78.2%	7.03 M
2	√	×	×	×	83.2%	47.6%	1.15 h	1.5 ms	77.8%	5.80 M
3	×	√	×	×	82.8%	47.4%	1.85 h	4.0 ms	85.1%	7.59 M
4	×	×	√	×	83.6%	47.7%	2.79 h	1.9 ms	89.6%	7.17 M
5	×	×	×	√	81.0%	47.1%	1.54 h	1.6 ms	81.5%	7.06 M
6	√	√	×	×	83.9%	48.2%	2.70 h	4.0 ms	82.9%	6.37 M
7	√	√	√	×	84.5%	48.1%	3.95 h	5.8 ms	90.5%	6.49 M
8	√	√	√	√	85.1%	48.6%	4.09 h	5.9 ms	91.1%	6.52 M

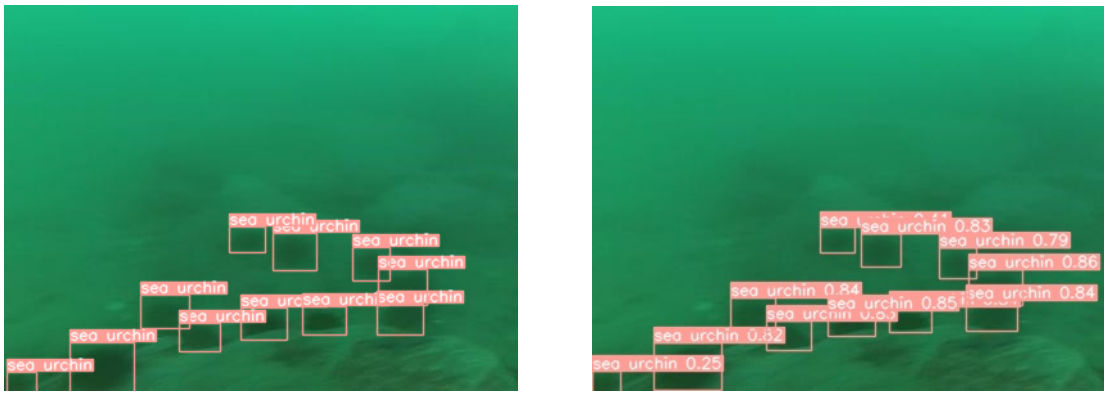
The results of mAP@0.5 and mAP@0.5:0.95 of the YOLOv5s baseline model were 80.8%, and 47.2%, respectively, and the number of parameters was 7.03 M. After using the improved CFnet structure, due to the optimization of the CFnet structure, the number of parameters of the model was reduced from 7.03 M to 5.80 M. At the same time, mAP@0.5 was also improved by 2.4%. After adding the Dyhead structure on the basis of CFnet, due to the efficient perception ability of the Dyhead structure, the mAP@0.5 value was further improved, reaching 83.9%. After adding the small object detection layer, the perception ability of the model for small objects was enhanced, making mAP@0.5 increase to 84.5%. Finally, the SE attention mechanism improved the model's generalization ability at the deepest layer of YOLOv5-CFDSDSE, and reached 85.1% for mAP@0.5 and 48.6% for mAP@0.5:0.95. The improvement of mAP@0.5 will also lead to the loss of other performance. While mAP@0.5 increased by 4.3%, the inference time was also increased to 5.9 ms. These data were derived when batch-size was 32. However, in the actual detection process, the inference time of 5.9 ms had little effect on the results, and the model can still obtain smooth detection results.

Detecting small underwater objects is one of the important purposes of YOLOv5-CFDSDSE. To compare the performance of each improved method in detecting small objects, this paper counted the percentage of the number of small objects detected by each improved method to the total number of small objects. The Small Object Detection Ratio (SODR) are shown in Table 2. It can be seen from Table 2 that the proposal of the SD layer was particularly significant for small objects, increasing the SODR from 78.2% to 89.6%. Combined with several other methods, the SODR in this data set was finally increased to 91.1%, which was 12.9% higher than that of YOLOv5s.

4.5. Visualization

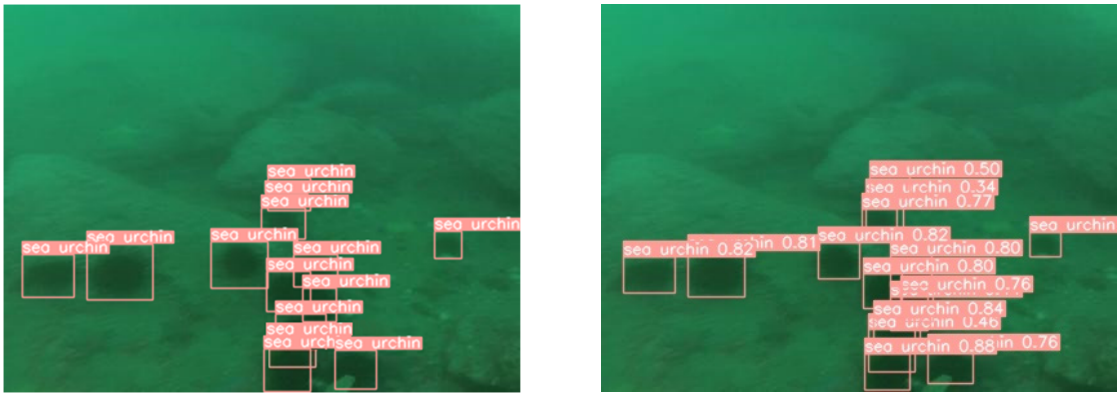
To verify the effect of the YOLOv5-CFDSDSE model in underwater image detection, typical underwater image scenes were selected from the dataset. The YOLOv5-CFDSDSE

model had a strong detection ability for dense small objects in complex underwater environments. The detection results of underwater fuzzy small objects are shown in Figure 8. The detection results of underwater dense small objects are shown in Figure 9. Comparing the detection results of the YOLOv5-CFDSDSE model with those of the YOLOv5s model, YOLOv5-CFDSDSE can identify targets that YOLOv5s cannot recognize. The comparison results are shown in Figures 10 and 11, where (b) is the YOLOv5s detection result, and (c) is the same image's detection result of YOLOv5-CFDSDSE.



(**a**) Ground Truth (**b**) YOLOv5-CFDSDSE detection results

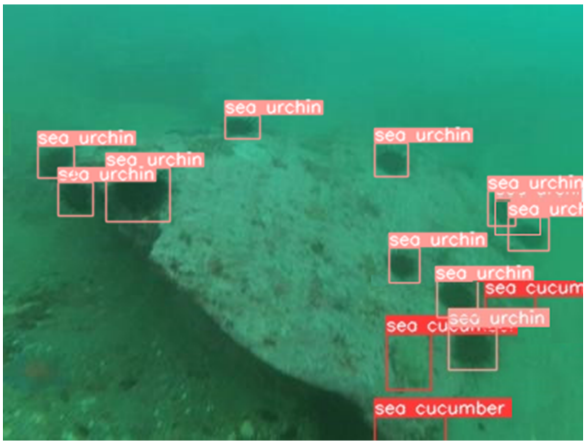
Figure 8. Underwater fuzzy object detection result diagram.



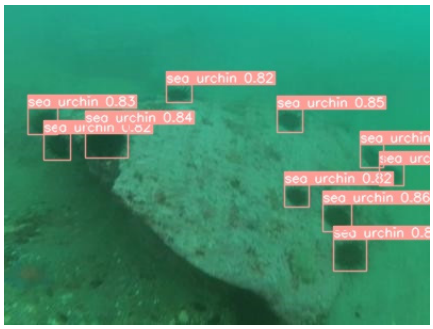
(**a**) Ground Truth (**b**) YOLOv5-CFDSDSE detection results

Figure 9. Underwater dense small object detection result diagram.

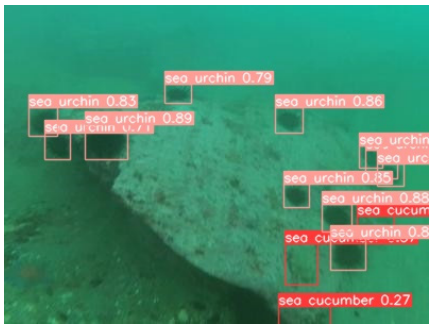
It can be seen from Figures 8 and 9 that YOLOv5-CFDSDSE had a good effect on object detection in blurred underwater scenes, could effectively extract features, and had excellent detection capabilities for dense small objects. As shown from (b) and (c) in Figure 10, YOLOv5s did not detect sea cucumbers, while the improved model YOLOv5-CFDSDSE successfully detected them, which are marked with a red box in (c). As shown from (b) and (c) in Figure 11, the improved model could detect sea urchins in a very low-definition state, while YOLOv5s could not detect them. After comparing the targets detected by the two models simultaneously, it was found that YOLOv5-CFDSDSE had higher accuracy for object detection in complex underwater environments. YOLOv5-CFDSDSE worked well for small object detection, but not all small objects could be detected. Objects below 1% of the image area were difficult to detect.



(**a**) Ground Truth

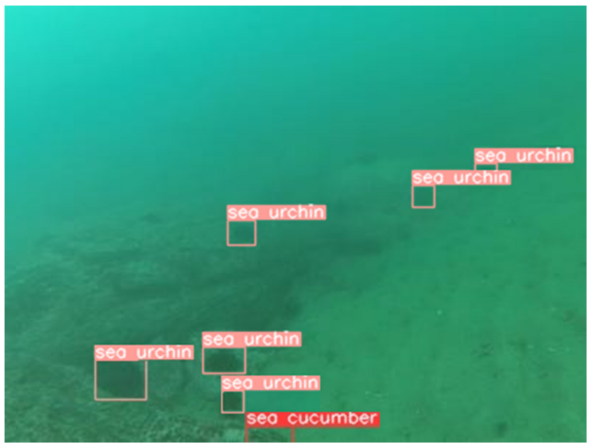


(**b**) YOLOv5 detection results



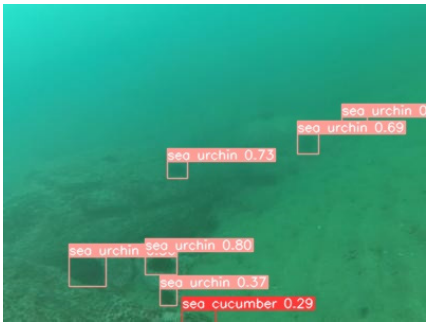
(**c**) YOLOv5-CFDSDSE detection results

Figure 10. Comparison diagram 1 of detection results.



(**a**) Ground Truth

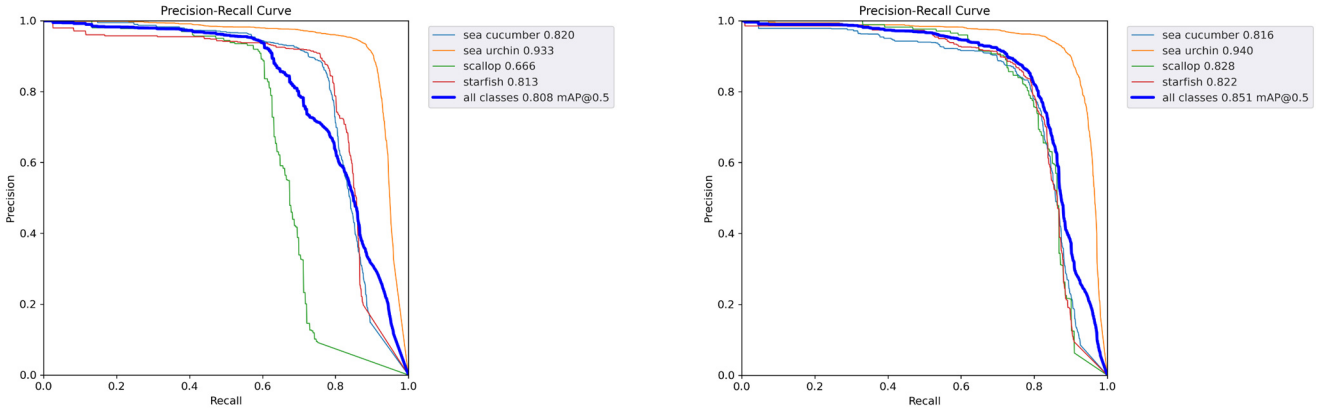
(**b**) YOLOv5 detection results



(**c**) YOLOv5-CFDSDSE detection results

Figure 11. Comparison diagram 2 of detection results.

Figure 12 shows the P-R diagram of the two models. The P-R diagram is a helpful tool in deep learning, which helped us understand the model's performance and made a trade-off between precision and recall.



(**a**) YOLOv5 (**b**) YOLOv5-CFDSDSE

Figure 12. Comparison of P-R diagram.

The shape of the PR curve can reflect the performance of the classification model. In general, the closer the PR curve was to the upper right corner, the better the performance of the classification model, as it had both higher precision and recall at this point. Comparing the P-R plots of the two models, we can find that the improved YOLOv5-CFDSDSE model performed better in Precision and Recall for object detection with the same dataset, and its accuracy was higher than YOLOv5s.

5. Conclusions

Underwater objects often appear small, dense, overlapping, and blurred, resulting in highly complex underwater object detection. This paper proposed a new model YOLOv5-FCDSTSE for underwater dense small object detection. It was based on the YOLOv5s model and improved C3 to CFnet structure, which reduced the number of model structure parameters while also improving the performance of detection. The multi-scale and multi-category object detection head Dyhead was used to replace the original Head, which enhanced the perception ability of the model. The SD layer was added to effectively combine feature information from different scales through multi-level feature fusion, which can retain more detailed information and improve the perception of small objects. Finally, an SE attention mechanism was added to the deep layer of the model to recalibrate each channel in the feature map. It placed more attention on the features that were useful for the current task. By learning the channel weight vector, the SE attention mechanism gave a larger weight to the channel with important information and a smaller weight to the channel with irrelevant information. The experimental results showed that the YOLOv5-CFDSDSE model achieved good performance on the self-made URPC_UODD underwater small object dataset.

The YOLOv5-CFDSDSE model proposed in this paper was significantly better than the baseline model in terms of accuracy and parameter quantity. Its detection frame rate could also reach 22FPS. However, to achieve better application results, the focus of the next step is to study further how to improve the detection frame rate technology. At the same time, because the detection effect of this model on occluded targets in some particularly complex underwater environments still needs to be improved, further research on how to solve the problem of occluded object detection is another direction of work in the future.

Author Contributions: Conceptualization, J.W. (Jingyang Wang); methodology, J.W. (Jingyang Wang) and Y.L. (Ying Li); software, Y.L. (Yujia Li); validation, J.W. (Junkai Wang); investigation, J.W. (Junkai Wang) and Y.L. (Yujia Li); writing—original draft preparation, Y.L. (Yujia Li) and J.W. (Junkai Wang); writing—review and editing, J.W. (Jingyang Wang) and Y.L. (Ying Li); visualization, Y.L. (Yujia Li) and J.W. (Junkai Wang); supervision, Y.L. (Ying Li) and J.W. (Jingyang Wang); project administration, Y.L. (Yujia Li) and J.W. (Junkai Wang); funding acquisition, J.W. (Jingyang Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation Foundation of Hebei Intelligent Internet of Things Technology Innovation Center under Grant AIOT2203, and by the Defense Industrial Technology Development Program under Grant JCKYS2022DC10.

Data Availability Statement: Data available on request due to restrictions privacy. The data presented in this study are available on request from the corresponding author.

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

References

1. Sun, K.; Cui, W.; Chen, C. Review of Underwater Sensing Technologies and Applications. *Sensors* **2021**, *11*, 7849. [CrossRef]
2. Pellegrino, F.A.; Vanzella, W.; Torre, V. Edge detection revisited. *IEEE Trans. Syst. Man Cybern. Syst.* **2004**, *34*, 1500–1518. [CrossRef]
3. Ehrenfried, K. Processing calibration-grid images using the hough transformation. *Meas. Sci. Technol.* **2002**, *13*, 975–983. [CrossRef]
4. Omachi, S.; Omachi, M. Fast Template Matching with Polynomials. *EEE Trans. Image Process.* **2007**, *16*, 2139–2149. [CrossRef]
5. Guenther, N.; Schonlau, M. Support Vector Machines. *Stata J.* **2016**, *16*, 917–937. [CrossRef]
6. Scornet, E. Random Forests and Kernel Methods. *IEEE Trans. Inf. Theory* **2016**, *62*, 1485–1500. [CrossRef]
7. Shichao, Z.; Xuelong, L.; Ming, Z.; Xiaofeng, Z.; Debo, C. Learning k for kNN Classification. *ACM Trans. Intell. Syst. Technol.* **2017**, *8*, 43.
8. 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.* **2017**, *39*, 1137–1149. [CrossRef]
9. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
11. Zhu, Z.; Liang, D.; Zhang, S.; Huang, X.; Li, B.; Hu, S. Traffic-Sign Detection and Classification in the Wild. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 26–1 July 2016.
12. Liu, K.; Peng, L.; Tang, S. Underwater Object Detection Using TC-YOLO with Attention Mechanisms. *Sensors* **2023**, *23*, 2567. [CrossRef] [PubMed]
13. Chen, Z.; Zhang, Z.; Dai, F.; Bu, Y.; Wang, H. Monocular Vision-Based Underwater Object Detection. *Sensors* **2017**, *17*, 1784. [CrossRef] [PubMed]
14. Chen, Y.; Ling, Y.; Zhang, L. Accurate Fish Detection under Marine Background Noise Based on the Retinex Enhancement Algorithm and CNN. *J. Mar. Sci. Eng.* **2022**, *10*, 878. [CrossRef]
15. Liu, J.; Liu, S.; Xu, S.; Zhou, C. Two-Stage Underwater Object Detection Network Using Swin Transformer. *IEEE Access.* **2022**, *10*, 117235–117247. [CrossRef]
16. Wei, X.; Yu, L.; Tian, S.; Feng, P. Underwater target detection with an attention mechanism and improved scale. *Multimed. Tools Appl.* **2021**, *80*, 33747–33761. [CrossRef]
17. Al Muksit, A.; Hasan, F.; Bhuiyan Emon, M.F.H.; Haque, M.R.; Anwary, A.R.; Shatabda, S. YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment. *Ecol. Inform.* **2022**, *72*, 101847. [CrossRef]
18. Zhang, C.; Zhang, G.; Li, H.; Liu, H.; Tan, J.; Xue, X. Underwater target detection algorithm based on improved YOLOv4 with SemiDSConv and FIoU loss function. *Front. Mar. Sci.* **2023**, *10*, 1153416. [CrossRef]
19. Shi, Y. An Underwater Target Wake Detection in Multi-Source Images Based on Improved YOLOv5. *IEEE Access.* **2023**, *11*, 31990–31996. [CrossRef]
20. Li, J.; Liu, C.; Lu, X.; Wu, B. CME-YOLOv5: An Efficient Object Detection Network for Densely Spaced Fish and Small Targets. *Water* **2022**, *14*, 2412. [CrossRef]
21. Li, W.; Zhang, Z.; Jin, B.; Yu, W. A Real-Time Fish Target Detection Algorithm Based on Improved YOLOv5. *J. Mar. Sci. Eng.* **2023**, *11*, 572. [CrossRef]
22. Wang, J.; Qi, S.; Wang, C.; Luo, J.; Wen, X.; Cao, R. B-YOLOX-S: A Lightweight Method for Underwater Object Detection Based on Data Augmentation and Multiscale Feature Fusion. *J. Mar. Sci. Eng.* **2022**, *10*, 1764. [CrossRef]
23. Ultralytics. yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 18 May 2020).

24. Chen, J.; Kao, S.-H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
25. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
26. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
27. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic ReLU. *arXiv* **2020**, arXiv:2003.10027II.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
29. Jiang, L.; Wang, Y.; Jia, Q.; Xu, S.; Liu, Y.; Fan, X.; Li, H.; Liu, R.; Xue, X.; Wang, R.; et al. Underwater Species Detection using Channel Sharpening Attention. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 20–24 October 2021; pp. 4259–4267.