Journal of *Marine Science and Engineering*

MDPI

Article

Lightweight Underwater Target Detection Algorithm Based on Dynamic Sampling Transformer and Knowledge-Distillation Optimization

Liang Chen *, Yuyi Yang, Zhenheng Wang, Jian Zhang, Shaowu Zhou and Lianghong Wu

School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China
* Correspondence: kentchen@hnust.edu.cn

Abstract: Underwater robot perception is a critical task. Due to the complex underwater environment and low quality of optical images, it is difficult to obtain accurate and stable target position information using traditional methods, making it unable to meet practical use requirements. The relatively low computing power of underwater robots prevents them from supporting real-time detection with complex model algorithms for deep learning. To resolve the above problems, a lightweight underwater target detection and recognition algorithm based on knowledge distillation optimization is proposed based on the YOLOv5-lite model. Firstly, a dynamic sampling Transformer module is proposed. After the feature matrix is sparsely sampled, the query matrix is dynamically shifted to achieve the purpose of targeted attention modeling. Additionally, the shared kernel parameter convolution is used to optimize the matrix encoding and simplify the forward-propagation memory overhead. Then, a distillation method with decoupled localization and recognition is designed in the model-training process. The ability to transfer the effective localization knowledge of the positive sample boxes is enhanced, which ensures that the model maintains the same number of parameters to improve the detection accuracy. Validated by real offshore underwater image data, the experimental results show that our method provides an improvement of 6.6% and 5.0% over both baseline networks with different complexity models under the statistical index of detection accuracy mAP, which also suggests 58.8% better efficiency than models such as the standard YOLOv5. Through a comparison with other mainstream single-stage networks, the effectiveness and sophistication of the proposed algorithm are validated.

Keywords: underwater target detection; Transformer; YOLOv5; lightweight; knowledge distillation

Citation: Chen, L.; Yang, Y.; Wang, Z.; Zhang, J.; Zhou, S.; Wu, L. Lightweight Underwater Target Detection Algorithm Based on Dynamic Sampling Transformer and Knowledge-Distillation Optimization. *J. Mar. Sci. Eng.* **2023**, *11*, 426. https://doi.org/10.3390/jmse11020426

Academic Editors: Dimitrios V. Lyridis and Charis Ntakolia

Received: 12 January 2023
Revised: 8 February 2023
Accepted: 13 February 2023
Published: 15 February 2023

1. Introduction

Underwater surveying is a crucial tasks in underwater salvaging, underwater rescue, and autonomous robot navigation operations. Underwater intelligent robots contain a large number of sensors, including sonar and optical cameras, to enable perception of the surrounding environment [1,2]. In recent years, with the further development of advanced vision-processing technology, low-cost optical cameras with integrated and efficient algorithms have shown more potential than sonar in achieving higher localization accuracy in high-resolution optical images, which have drawn extensive research interest [3,4].

Due to their powerful feature-extraction ability, existing deep-learning-based methods for underwater image target detection use convolutional neural networks (CNNs) [5,6]. However, targets in underwater vision analysis tasks are influenced by the environment, making it difficult to obtain accurate localization information [7–9]. Deep learning-based methods learn high-dimensional features of images through large amounts of data, which could alleviate the problem of the difficulty localizing underwater targets.

Existing detection methods can be divided into two-stage detection and single-stage detection network frameworks. In the two-stage network, potential target areas are first

extracted and then the localization boxes are fine-tuned with global information to improve detection accuracy [10–12]. Single-stage detection algorithms use a backbone to extract features and multi-scale fusion through feature pyramids [13–15]. The structural paradigm of direct inference makes the model simple and efficient, and it is widely used in most underwater intelligent robots. However, this type of algorithm lacks attention to global environmental information, resulting in poor localization accuracy of underwater targets. Current single-stage algorithms in atmospheric environments often add a spatial attention structure to the multi-scale fusion structure to alleviate the lack of global information [16,17]. In the literature [18], ROI fusion is performed on high-level feature maps to deal with the problem of overlapping targets underwater. The authors of [19] propose a channel-attention module combined with a sharpening preprocessing method to improve underwater target detection accuracy. However, the hardware requirements of the algorithm limit its applicability in underwater mobile robots, especially for real-time purposes. The model with the above improvements is more complex, which is not suitable for the deployment of underwater robotic devices.

To balance the performance and efficiency of the model, existing research has focused on half-precision data, model pruning, and knowledge distillation methods for processing large models while maintaining good accuracy and minimizing the resources required. Geoffrey Hinton and other researchers introduced the knowledge-distillation (KD) method. The current KD method in target detection is mainly divided into response-based distillation and deep-feature-layer-based distillation. Response-based distillation transfers recognition knowledge using a logit simulation probability distribution to approximate the teacher model manifold [20]. However, this method only transfers classification knowledge and lacks feature learning under spatial constraints, which could lead to low knowledge-transfer efficiency, making it difficult to effectively improve detection accuracy. Knowledge distillation based on deep features is biased to enforce the consistency of the deep features [21,22], but it is difficult to separate which knowledge is beneficial for detection and which is beneficial for recognition. Knowledge distillation in underwater target detection has only been studied in a small number of related aspects, focusing mainly on the consistent enhancement of the deep feature layer [23]. The above algorithm reduces the cost of constructing global information features by downsampling when training small student models, resulting in a global feature-point-fusion structure for student models that lack effective feature sampling. Additionally, the use of feature-space-consistency distillation methods introduces redundant information or even information that should be suppressed [24], making it difficult to guarantee the detection accuracy of the algorithm.

To solve the above problems, this paper presents a dynamic sampling Transformer module applied to deep feature fusion based on YOLOv5-lite, which is used to compensate the global attention capability of the network. To achieve a lightweight design of the model, the module downsamples all three mapping matrices, Q, K, and V, to reduce the memory consumption of forward feature propagation. The extraction of key information is enhanced by dynamic sampling of the query matrix, which balances the detection accuracy and detection efficiency of the algorithm. In addition, a grid-localization distillation strategy with decoupled localization and recognition knowledge is proposed in the model KD process. Unlike the conventional KD process, the algorithm in this paper filters the localization information along the grids, which is more in line with the training process of YOLO, simplifies the training cost of the algorithm, and ultimately improves the detection accuracy of the algorithm. The whole process is shown in Figure 1.

The contributions of this paper are mainly summarized as follows.

(1) A lightweight Transformer module with dynamic sampling is proposed, revealing the importance of the Transformer sparse sampling dynamic transformation in an underwater environment.
(2) A knowledge-distillation framework for decoupling localization and recognition information applicable to YOLO is integrated, and the effectiveness of this detection algorithm is analyzed.

(3) The training cost of decoupled distillation of localization and recognition information is investigated by experimental comparison, and the effectiveness of this paper's algorithm is verified by comparison with other mainstream detection algorithms.

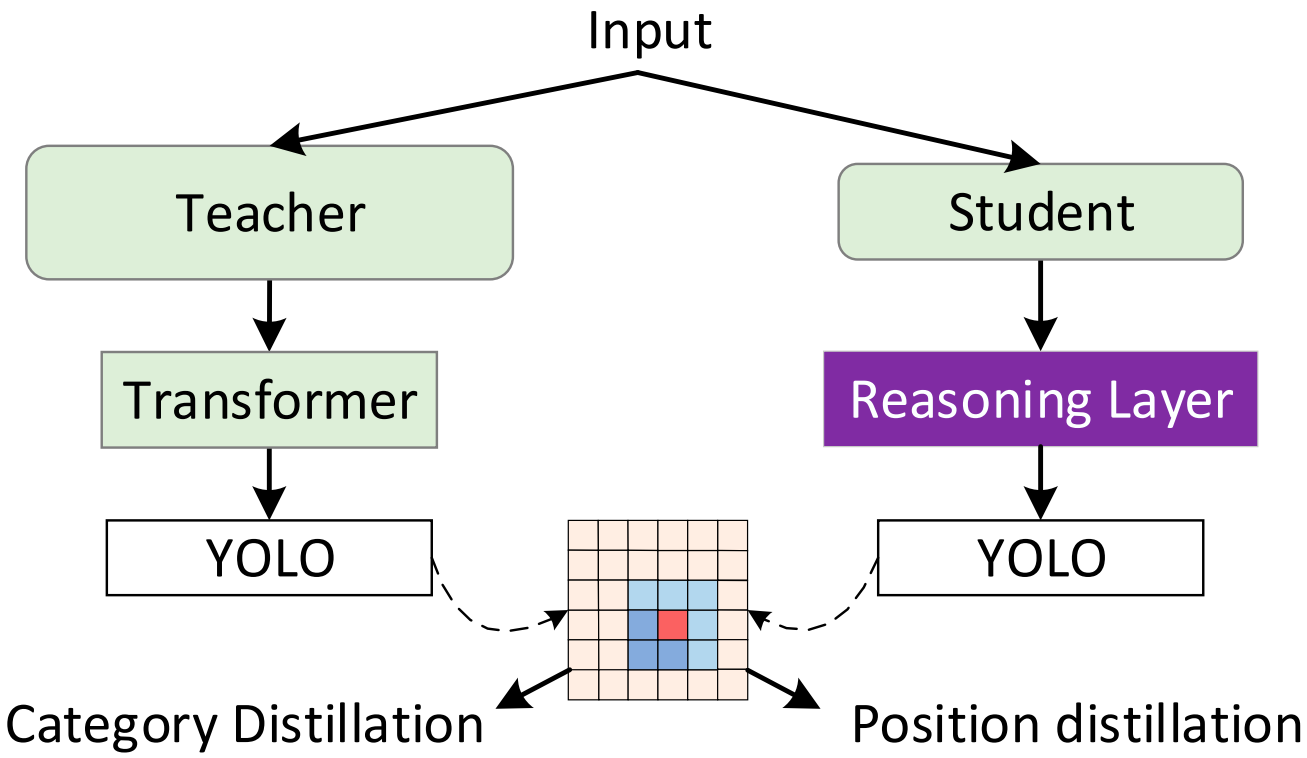

Figure 1. The overall architecture of our lightweight underwater target detection.

The rest of this paper is organized as follows. In Section 1, the technical lineage is sorted out, and the problems to be solved are introduced. Related work on target detection and knowledge distillation is presented in Section 2. The model improvement and knowledge distillation methods are introduced in Section 3. This is followed by a comprehensive discussion in Section 4. The conclusion is drawn in Section 5.

2. Related Work

2.1. Target Detection Models

To accommodate real-time operations, single-stage target detectors are gradually being adopted in underwater application deployments, where Anchor-free networks eliminate the computation of post-processing, bringing good detection performance [15]. The literature [25,26] applies a lightweight backbone to improve the Anchor-free CenterNet to further enhance the inference speed. Information exchange is enhanced through multi-scale feature fusion, and Vision Transform (ViT) is applied to fully focus and mine the information on holothurian ecological scenarios of different scales and spaces. However, the performance of the Anchor-free detector gradually degrades when the underwater environment tends to be complex, and the targets overlap more. The Anchor-based detector achieves stability by adapting to different morphological targets with pre-set prior box parameters. Reference [27] applied an attention mechanism to detect underwater targets on the basis of Anchor-free SSD. The authors of [28] applied GAN networks to enhance the data distribution before detection. However, the above methods are more complex and cannot simplify the model to improve detection. In the most recent method, the authors of [29] introduced an efficient framework that combines the classic effective design of backbone and FPN. The authors of [30–32] detected underwater targets using a single-stage detection algorithm through well-fused features.

2.2. YOLO-Lite Detection Algorithms

YOLO, a one-stage Anchor-based network model widely used in industry, has been tested in most application scenarios. YOLO-lite uses the computationally intensive convolutional shuffle module to reduce model parameters, reduce memory throughput overhead, and improve network inference efficiency. In its feature-fusion structure (neck), the multi-scale feature maps are fused by bipartite paths, and each image region is processed separately using the convolutional unit, which lacks the potential relationships between feature regions. The authors of [33] combined FPN-attention on the basis of YOLOv4-tiny to enhance the tight connection between extracted features. However, the number of parameters for building feature attention matrices is constantly complicated by structural improvements, and their accuracy and efficiency are difficult to balance. Subsequent researchers have gradually combined detection algorithms with knowledge distillation strategies.

2.3. Knowledge Distillation Strategy

Knowledge distillation is implemented through the knowledge migration from the teacher model to the student model. The algorithm flow is shown in Figure 2 by attaching the feature relationships already learned by the teacher network to the student network during training. In this way, the detection accuracy of the student model is improved, and the student model can be widely used in the compression model.

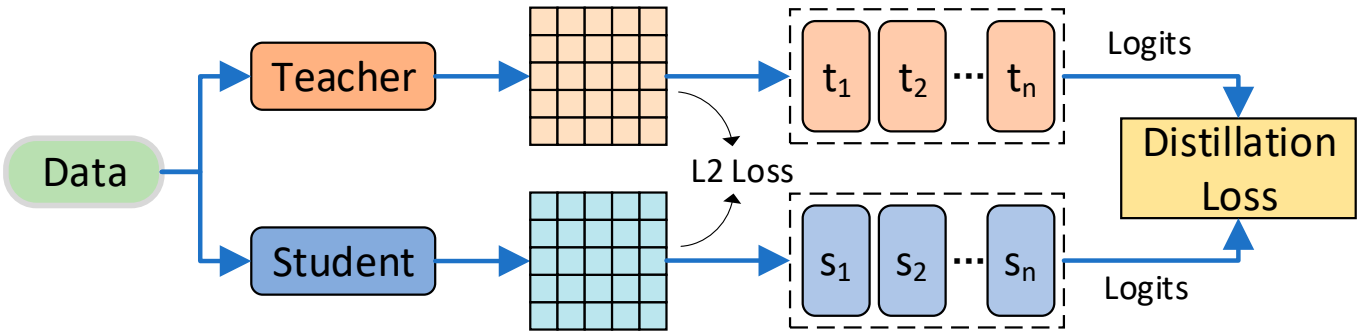

Figure 2. Knowledge distillation structure diagram.

Knowledge distillation is trained by the teacher network by guiding the student network, and the output categorical information is transformed into soft logits, which converge the knowledge distribution of the student network to that of the teacher network through a KL scatter. Common knowledge distillation methods are inefficient due to primitive logit mimicking techniques, as they transfer only categorical knowledge and ignore the significance under spatial information constraints. Reference [34] uses L2 Loss to directly transfer the knowledge of the feature map containing localization and category information by approximating the elements of the high-level feature map, which would introduce negative knowledge in the teacher's network, influencing the detection accuracy of the student's network. The recent methods are shown in Table 1, including that of [35,36], which designs intermediate conversion modules for feature layer distillation. Reference [37] uses semantic information on channel-dimension distillation, and they both use feature-based distillation methods.

Table 1. A summary of feature-based distillation methods.

Methods	Knowledge Types	Knowledge Sources
DeFeat [34]	FPN Features	Hint layer
RDM [35]	Prototype Generation Module Features	TS-Space
PGD [36]	Key Predictive Regions	Hint layer
CWD [37]	Channel Distribution	Hint layer

In the YOLO detector (head), each branch detector outputs localization and identification information within a grid region, simplifying the detection process. However, it also consequently restricts mutual information to a single branch path, limiting the algorithm's performance. Multiple branching paths have been shown to reduce detection efficiency [38]. In contrast, KD can deliver both localization knowledge and recognition knowledge of the teacher network in the detector, bringing performance gains to the detection algorithm. However, none of the existing KD algorithms consider the detection process of assigning a positive sample Anchor in the YOLO network to deliver the localization knowledge and recognition knowledge uniformly, which could cause an increase the risk of overfitting during the training process of YOLO, leading to a decrease in detection accuracy due to the inability to accurately locate the target.

3. Proposed Method

YOLOv5-lite uses the depth parameter to control the number of model parameters in different interval ranges. In this paper, we mainly introduce dynamic sampling Transformer on YOLOv5s-lite to enhance the network's ability to establish global information association in underwater target detection. In addition, localization knowledge distillation is used to alleviate the lite model to manage the problem of blurred underwater target coordinate boundaries in complex environments, as well as to improve the generalization capability of the model.

3.1. Dynamic Sampling Transformer

In order to effectively aggregate the features of image regions, this paper constructs a Reasoning Layer at the top layer of the feature fusion layer [16]. Additionally, a Transformer structure is proposed to build a random residual structure together with convolutional branching, which is beneficial for compensating the performance limitation of local attention. The structure of the proposed algorithm in this paper is shown in Figure 3.

The reasoning layer is built on top of YOLO's neck structure and is used to build the information association of the fused features. To achieve this operation, the pixels are encoded, and then the similar information in the neighborhood is extracted by using the channel-separable convolution through the local information aggregation module (Local Aggregation) and aggregated by using the pooling operation. In order to obtain an inter-domain multi-scale view, a larger perceptual field in the neighborhood is obtained by designing the 3×3 grouped convolution of separable convolution as an expanded convolution with shared kernel parameters without increasing the number of parameters. It can be expressed as follows.

$$F_{\text{Aggregation}} = WConv(\sum_{r \in \mathbb{R}} SiLU(DConv_r(F_{in}, W_A^{k=3}, g = d))) \tag{1}$$

where F_{in} is the encoded feature map, $SiLU(\cdot)$ is the activation function, $DConv_r(\cdot)$ is the expansion convolution operation with r as the expansion rate, $W_A^{k=3} \in \mathbb{R}^{k^2 \times d \times d/g}$ is the kernel weight, and g is the number of grouped convolution sets. The multi-scale spatial perception of the neighborhood can be obtained by setting different expansion rates to improve the local aggregation ability. The multi-scale spatial perception of the neighborhood can be obtained by setting different expansion rates to improve the local aggregation ability. Finally, the dimensional mapping is performed using $WConv(\cdot)$ pointwise convolution to obtain the representative feature point information, which is used as the input of the subsequent multi-headed self-attention. The formula of multi-headed self-attention is shown below.

$$Z^{(m)} = SA(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, m = 1, \cdots\cdots, N \tag{2}$$

$$Q = xW_q, K = xW_k, V = xW_v \tag{3}$$

$$Z = Concat(Z^{(1)}, \cdots\cdots, Z^{(N)})W_o \tag{4}$$

where Q and K are used as the query matrix and the key matrix obtained by linear projection of the aggregated features $X \in \mathbb{R}^{n \times d_m}$, which is used to model the feature relationships, where $n = w \times h$, d_m is the feature dimension. The computational complexity of performing encoding and the matrix dot product similarity calculations is $O(2d_m n^2 + 4d_m^2 n)$, which is proportional to the space and dimension squared and occupies a large amount of forward-propagation memory space. The current compression approach is to downsample the eigenspace of the key matrix K and the value matrix V. A large number of parameters for obtaining global information are still retained when encoding the query matrix Q. While encoding each element is non-essential [39], sparse sampling with r as the compression rate can balance performance and efficiency. Therefore, the text downsamples the matrix when computing the self-attention, compresses the feature matrix representative information to make the representative feature points sparser, and upsamples the spatial information to reconstruct after the attention computation is completed. Finally, the output of each attention head is channel-concatenated, and the final result is obtained by linear projection of the output mapping matrix.

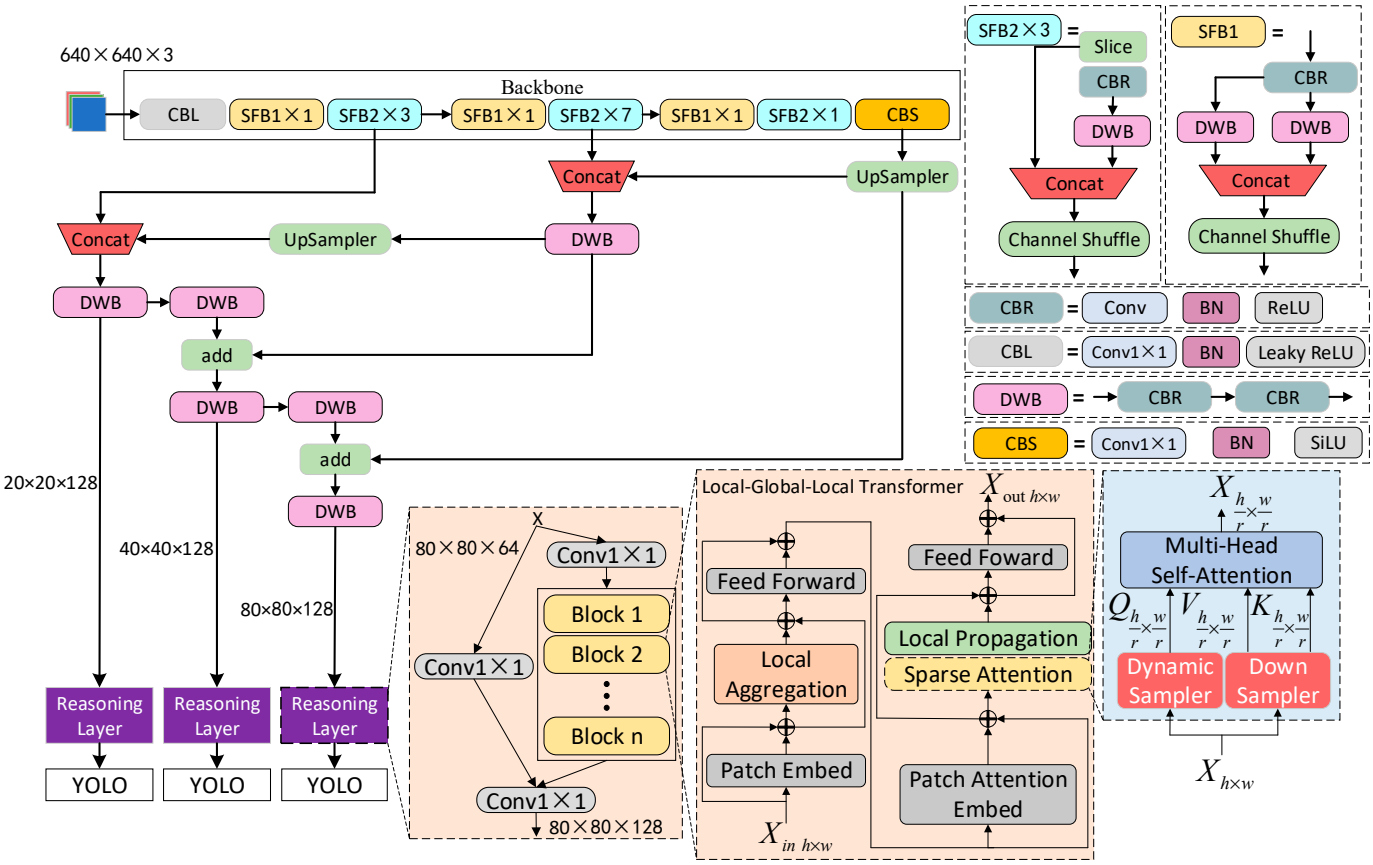

Figure 3. The YOLO model's Reasoning Layer improvement structure diagram.

To reduce the number of parameters for forward propagation, uniformly sparse sampling of representative feature information is performed as the K and V matrices of multi-headed attention. The key matrix K modeled for the global feature map for attention is ensured, and the computational complexity of this part is $O(r^2 n d_m)$. However, uniform sampling cannot dynamically obtain information richer in target features, so a position-offset module is designed by sampling points of the query matrix Q. The sampled point offsets are extracted using grouped convolution, and the process of position decoding in the literature [40] is referenced. By adding the offset with the reference sampling points and

then bilinear interpolation, the position-sampling points of the query Q with representative information are finally obtained, and the offset structure is shown in Figure 4.

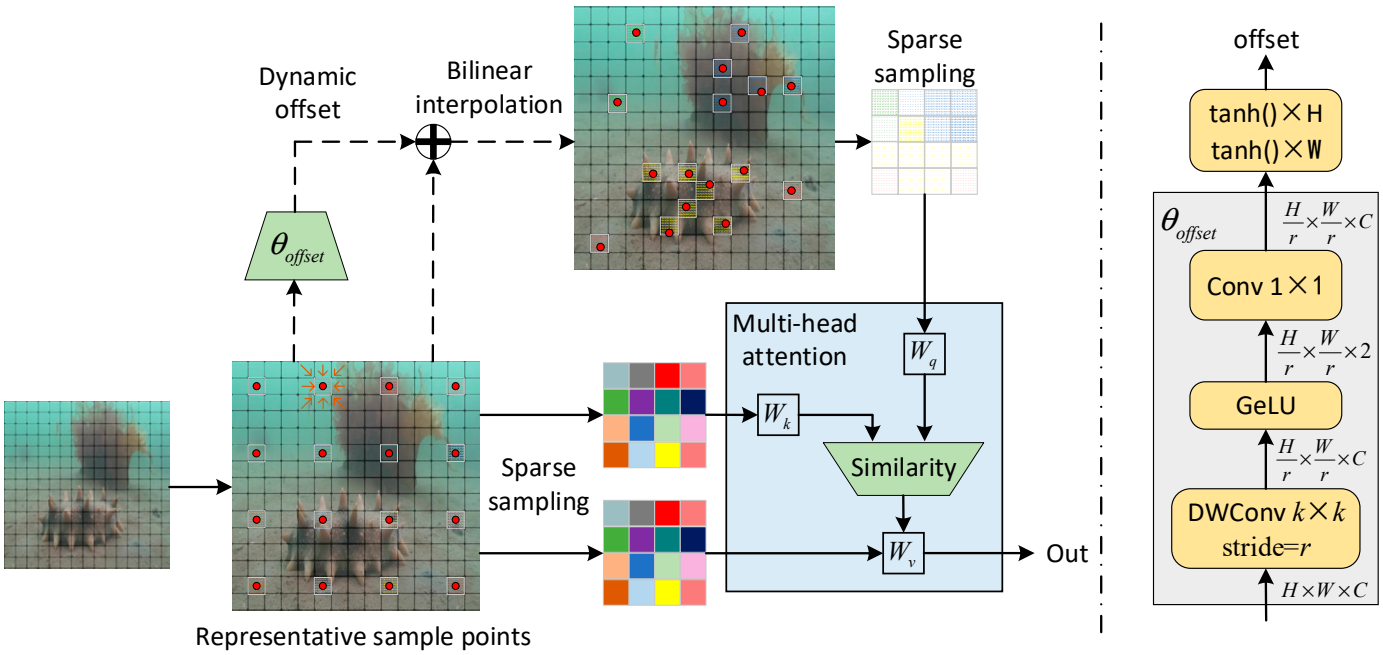

Figure 4. Dynamic sampling Transformer self-attention structure diagram.

As depicted in Figure 4, the sampling position offset module consists of the right region. By dynamically sampling the feature map of aggregated representative information, the valuable query point elements can be extracted in a targeted manner. The neighborhood centroid is used as the reference point and the offset $\theta_{offset}(\cdot)$ is obtained using the lightweight network branch $\Delta p = \theta_{offset}(p)$, $\Delta p \in \mathbb{R}^{\frac{H}{r}\times\frac{W}{r}\times 2}$. Therefore, the dynamic sampling of sparse feature points can be expressed as follows.

$$\Theta(X;(p_x,p_y)) = \sum_{(r_x,r_y)} S(p_x,r_x)S(p_y,r_y)X[r_y,r_x,:] \tag{5}$$

$$\hat{x}_d = \Theta(X;p+\Delta p) \tag{6}$$

where $\Theta(\cdot;\cdot)$ is used to implement bilinear binary interpolation to calculate the coordinate element values of dynamically sampled points. p_x, p_y denote arbitrary (fractional) position coordinates. $X[r_y,r_x,:]$ is the feature value of sparse sampling by index point and $S(a,b) = \max(0,1-|a-b|)$. $\hat{x}_d$ is the feature point for dynamic sampling. It is worth noting that this paper uses the spatial attention module to encode the location of the features before sampling dynamically, which can highlight representative aggregated information, so the calculation of multi-headed attention is shown below.

$$\hat{Q} = \hat{x}_d W_q, \hat{K} = \hat{x} W_k, \hat{V} = \hat{x} W_v \tag{7}$$

$$Z^{(n)} = Softmax(\frac{\hat{Q}^{(n)}\hat{K}^{(n)T}}{\sqrt{d}})\hat{V}^{(n)} \tag{8}$$

where $\hat{x}$ is a uniformly sparsely sampled feature point for obtaining the global view. $Z^{(n)}$ is the output of the n-dimensional self-attentive module. Finally, the feature maps are upsampled in the local location propagation module, and the attention feature maps are reconstructed in r-expansion multiples using transposed convolution. Propagating the attention relation of the sampled points in the neighborhood makes the module establish

the local connection of the feature map at a lower cost and improves the network detection performance.

3.2. Positioning Distillation

Lightweight networks cannot adapt to the complex underwater environment and require more data for training. This paper suggests a distillation method that separates localization information from recognition information to transfer the more effective positive sample-localization knowledge. Unlike the consistent propagation of deep feature maps, the localization and recognition knowledge are transmitted separately. The recognition knowledge is applied to KL scatter-learning teacher network probability distribution, which can be represented as follows.

$$f_{cls}^{comb}(c_i^{gt}, \hat{p}_i, p_i^T, \hat{o}_i^T) = f_{cls}^{CE}(p_i^{gt}, \hat{p}_i) + \lambda_1 \cdot \hat{o}_i^T \cdot f_{cls}^{distill}(p_i^T, \hat{p}_i) \tag{9}$$

$$f_{cls}^{CE}(p_i^{gt}, \hat{p}_i) = -\frac{1}{N}\sum_i \sum_{c=1}^{M} p_i^{gt} \log(\hat{p}_i) \tag{10}$$

$$f_{cls}^{distill}(p_i^T, \hat{p}_i) = \sum_i Te(\hat{p}_i) \log \frac{Te(p_i^T)}{Te(\hat{p}_i)} \tag{11}$$

where $c_i^{gt}, \hat{p}_i, p_i^T, \hat{o}_i^T$ denote the true category label, student network prediction category, teacher network prediction category, and teacher network prediction target confidence, respectively. $Te(x) = \frac{e^{x_i/T}}{\sum_k^N e^{x_k/T}}$ denotes the conversion of the prediction output to soft logit by distillation temperature T. Usually, the category distillation loss allows the model to be fitted to the teacher model through the KL scatter, and the cross-entropy loss (CE Loss) is utilized as the loss function for the multi-category task. Due to the difficulty of underwater image data acquisition, the sparsity of some target categories or the presence of labeling noise leads to the susceptibility of difficult categories. It is difficult for the CE Loss used in YOLOv5-lite to fit such data distribution. Therefore, this paper introduces a difficult-sample balancing factor to improve the detection of difficult samples [14]. The computational formula and its Taylor expansion can be expressed as follows.

$$f_{cls}^{FL} = -(1-\hat{p}_i)^{\gamma} \log(\hat{p}_i) = \sum_{j=1}^{\infty} \frac{1}{j}(1-\hat{p}_i)^{j+\gamma} \tag{12}$$

In this case, γ is used to balance the imbalance problem of difficult and easy samples. This term reduces the tendency of CE Loss to bias most categories, but it is difficult for this modulation parameter to target and adjust to different data. The derivation of the above equation shows that the gradient in backpropagation is concentrated in the leading polynomial [41] and provides a constant gradient that makes the loss emphasize most classes. Therefore, perturbing the important leading polynomial coefficients can improve the robustness of the system, while adjusting the first polynomial maximizes the enhancement effect. This can be expressed as follows.

$$f_{cls}^{Poly} = -(1-\hat{p}_i)^{\gamma} \log(\hat{p}_i) + \varepsilon(1-\hat{p}_i)^{\gamma+1} \tag{13}$$

where ε is the modulation factor of an important polynomial, and when $\varepsilon > 0$, it can improve the accuracy of unbalanced data. Therefore, the polynomial modulation factor is added to Focal Loss to replace the original CE Loss, so that the network optimization can be adaptive to the imbalance of the underwater target class.

The localization task, as an important part of the detection algorithm, should be distinguished from the classification knowledge distillation, so this paper provides targeted optimization of the location prediction loss. The localization knowledge is used to determine the target information within the enclosed Anchor position, and the target confidence

is introduced in the distillation information. In addition, YOLO divides the network feature map into grids for target identification, and the target center grid contains the most accurate position information. The preset Anchor box in the adjacent grid also covers a certain range of target areas, which can also contribute knowledge to the target location determination. Therefore, the value-region screening is performed in the neighboring grids to provide more localization knowledge of the target positive sample Anchor regression, and the localization knowledge algorithm flow proposed in this paper is shown in Figure 5.

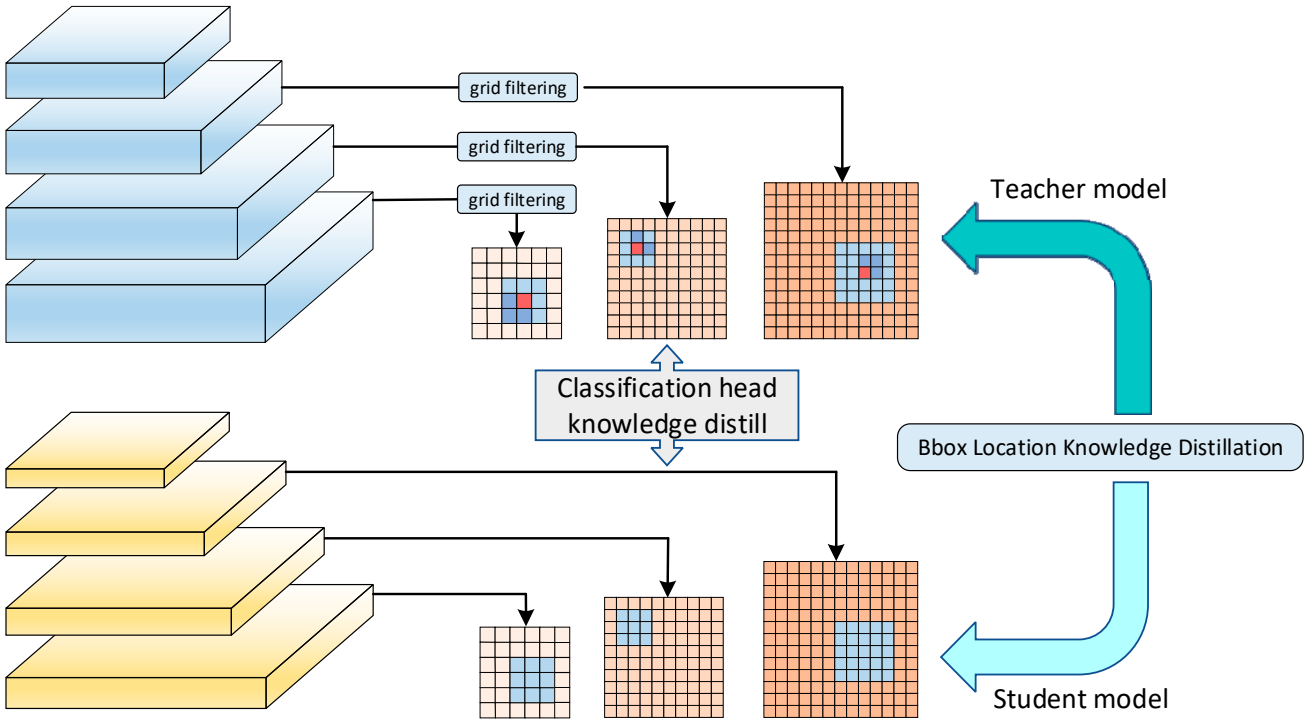

Figure 5. Model distillation flow chart.

The teacher model is first given to provide valid knowledge of the adjustment parameters of the N bounding boxes predicted at each element position in the last layer of the network, which includes the target confidence and category information. Thus, its output dimension is $N \times (C + 5)$, where C denotes the category index. Thereafter, the localization output is separated, and the valid Anchor location index I in the surrounding grid is determined by comparing it with the real label in the neighboring grid, and I $\times$ N valid location information knowledge is passed to the student network as a positive sample. Finally, since the candidate boxes in the center grid have higher confidence than the candidate boxes in the domain, the distance scaling factor between the grids is weighted on the confidence to maintain this predictive value. The target confidence is also passed as the target information weight of the enclosing box together with the location information. The coordinate regression loss can be expressed as follows:

$$f_{bbox}^{comb}(b_i^{gt}, \hat{b}_i, b_i^T, \hat{o}_i^T) = f_{bbox}^{MSE}(b_i^{gt}, \hat{b}_i) + \lambda_2 \cdot \hat{o}_i^T \cdot f_{bbox}^{distill}(\delta(b_i^T), \hat{b}_i) \tag{14}$$

where $b_i^{gt}, \hat{b}_i, b_i^T$ is the labeling label of the target enclosing boxes, the Anchor adjustment parameter output by the student network, and the Anchor adjustment parameter output by the teacher network, respectively. $\delta(\cdot)$ is the effective grid filter, and the grid-filtering algorithm can be expressed as Algorithm 1.

Algorithm 1 The grid-filtering algorithm

Input: A set of anchor boxes $B_f^a = \left\{B_{i_f}^{\alpha}\right\}$ and a set of ground truth boxes $B^{GT} = \{B_f^a\}$. Positive threshold β_{pos} of label assignment. H_f and W_f is the size of f-th Feature layer output. J is the number of ground truth boxes.
Output: $A_f = \{a_{i_f j}\}$, $a_{i_f j} \in [0, J \times n \times m]$ encodes final position index of the grids, where n is the number of pre-set anchor boxes and m represents the m fields of the center grid.
1: Compute $\max\left(\frac{anchor_W}{W_{GT}}, \frac{W_{GT}}{anchor_W}\right)$ and $max\left(\frac{anchor_H}{H_{GT}}, \frac{H_{GT}}{anchor_H}\right)$
2: $G_{vl} = \gamma_{distance} G_{center}$
3: Filter the positioning grid with indices of G_{vl}
4: **return** $indices_{grid}$

Firstly, the Anchor candidate boxes that conform to the real annotation boxes are obtained by thresholding the aspect ratio between the Anchor boxes and the real annotation boxes at J kinds of resolutions. Additionally, the valuable grid boxes in the m-neighborhood outside the centroid are obtained by the proportional size. Then, the distance weighting is performed within the neighborhood grid. Knowledge transfer to the student network is performed within the centroid grid and the valuable neighborhood range grid. The effective grid filtering is performed to reduce redundant localization information or even harmful localization information. Thus, the final loss function can be expressed as follows.

$$L_{all} = f_{cls}^{Poly}(p_i^{gt}, \hat{p}_i) + \lambda_1 \cdot \hat{o}_i^T \cdot f_{cls}^{distill}(p_i^T, \hat{p}_i) + f_{bbox}^{comb}(b_i^{gt}, \hat{b}_i, b_i^T, \hat{o}_i^T) + f_{obj}^{BCE}(o_i^{gt}, \hat{o}_i) + \lambda_3 \cdot \hat{o}_i^T \cdot f_{obj}^{distill}(o_i^T, \hat{o}_i) \quad (15)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the loss balance factors to balance the weight of knowledge distillation for different prediction tasks; $f_{obj}^{BCE}(\cdot)$ is the binary cross-entropy loss function of target confidence; and $f_{obj}^{distill}(\cdot)$ is the target confidence distillation loss function, which can be regarded as a regression task, so the regression is performed with mean squared loss. By decoupling localization knowledge and identification knowledge, targeted localization knowledge distillation is performed according to YOLO's Anchor matching strategy. The forward polynomial gradient weights of the classification loss function are boosted so that the algorithm detection and recognition capabilities are improved.

4. Experimental Verification and Analysis

4.1. Experimental Dataset

The proposed algorithm is investigated in optical vision-target detection of underwater operational equipment. To improve the detection accuracy of the algorithm while ensuring the light weight of the model, we therefore use 2900 real offshore underwater image datasets from the underwater robot professional contest (URPC) (https://www.curpc.com.cn/ (accessed on 1 September 2020) to divide the training set, validation set, and test set by 7:2:1 for training, which includes different environments such as reefs, mud and sand, gullies, water plants, etc. The data are images taken in shallow waters of the Bohai Sea without artificial light. The target scale is widely distributed, which can provide generalized validation for the algorithm application. This dataset is a dynamic image taken by an underwater operational robot during underwater seafood recovery in Bohai Bay, which was used in an underwater operational robot target-detection competition.

4.2. Implementation Details

In this paper, the training environment was the Ubuntu system, and the graphics card RTX2080-11G was used as the training device to transfer the COCO pre-trained network model, the batch size was set to 16, the image size was 416 × 416, and 500 rounds of training were iterated. For more effective training, a cosine learning rate optimization strategy was used to set the label smoothing parameter 0.005, and $\lambda_1 = \lambda_2 = \lambda_3 = 0.95$. To reduce the memory consumption of forward propagation, the compression rates of the feature layers of the three detector inputs were set to be 8, 4, and 2. The number of class instances, real

frame visualization, real frame label centroids, and aspect ratio scatter plots of the dataset are shown in Figure 6.

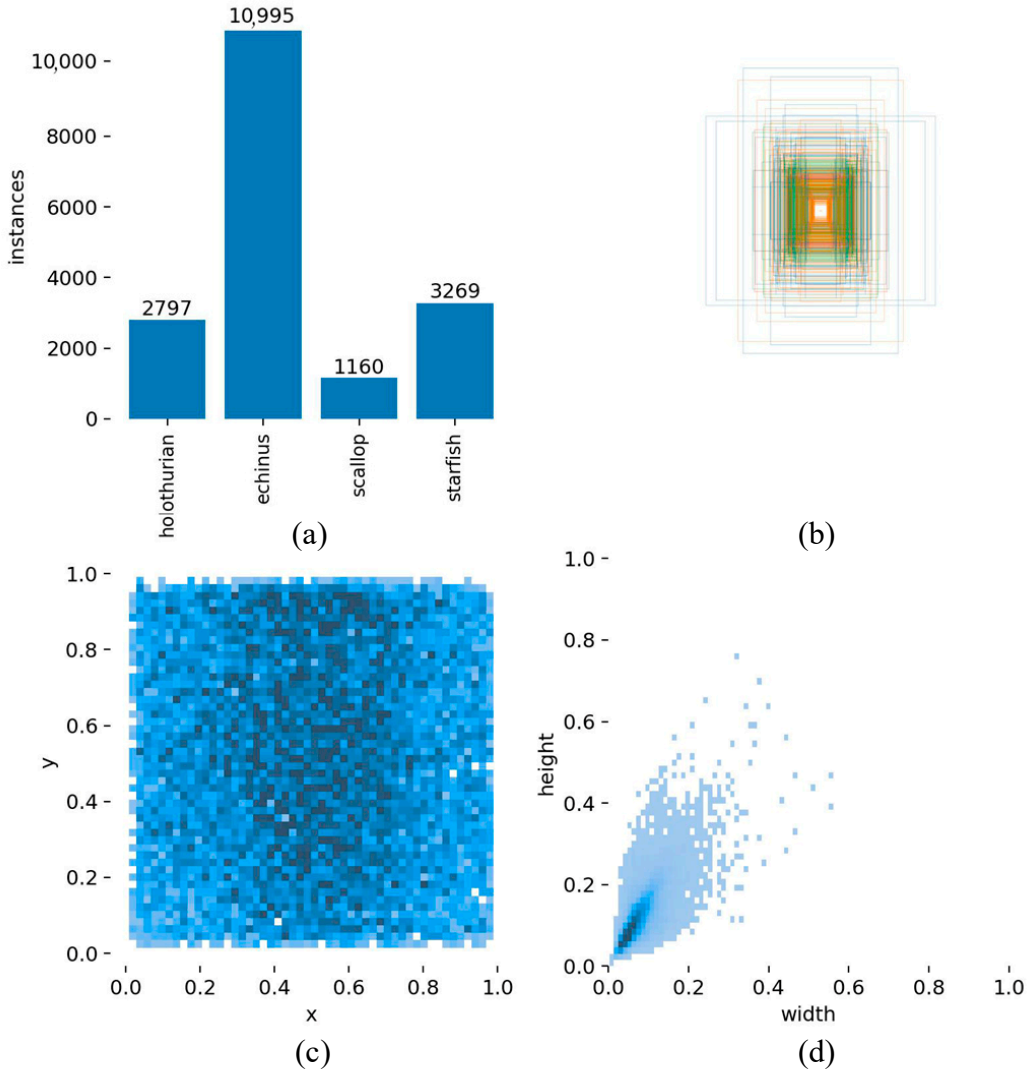

Figure 6. Statistical chart of dataset attributes. (**a**) is a count of the number of instances in the dataset. (**b**) is the target box visualization. (**c**) is the target box centroid location statistics. (**d**) is the target box aspect ratio scatter plot.

Figure 6a shows the instance labeling statistics, which represents the statistics of the number of different categories of instances in the dataset. Figure 6b shows the visualization of the unified centroids of the positioned labeling boxes. Figure 6c shows the normalized location distribution of the centroids of the labeled boxes, and Figure 6d shows the normalized distribution of the length and width of the labeled boxes. It can be seen that the experimental dataset suffers from unbalanced instance categories, with the minimum category is nearly 1/10 of the maximum category. Additionally, the centroids are primarily based on the uniform distribution of image centers with unbalanced aspect ratios and various scales of annotation boxes.

4.3. Dynamic Sampling Branch Analysis

To better illustrate the internal variation of the dynamically sampled Transformer, the Fourier transform of the internal features of the multi-headed attention module was analyzed, as shown in Figure 7.

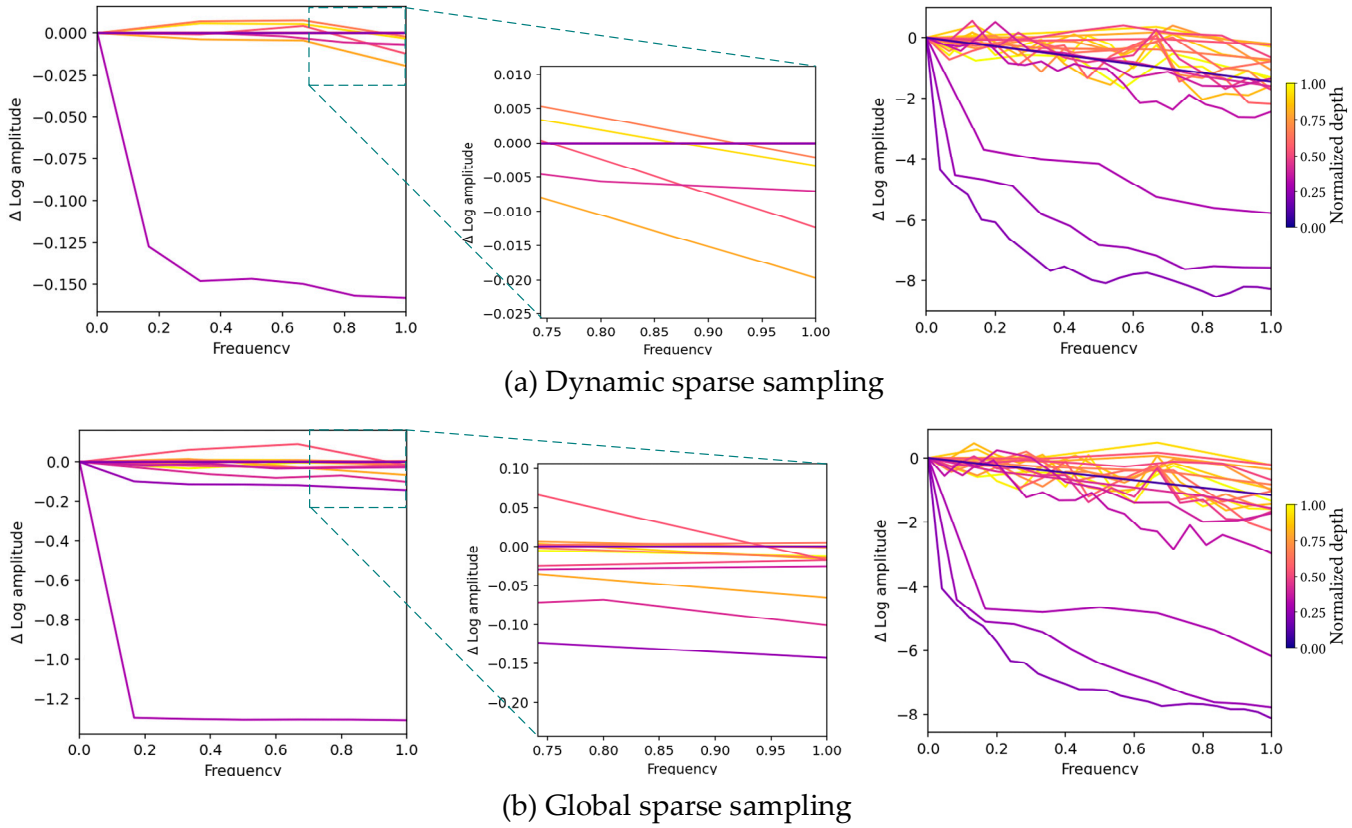

Figure 7. Fourier transform diagram of intermediate features.

The self-attention was calculated after feature aggregation and dynamic sampling with the three-level feature layer of YOLO, as shown in row (a). Δ Log amplitude of high-frequency signals is the difference between the log amplitude at the normalized frequency 0.0π (the center) and at 1.0π (the boundary). For better visualization, we only provide the half-diagonal components of two-dimensional Fourier-transformed feature maps. The left subplot of Figure 7a shows the Fourier transform results after sparse downsampling of the self-attention module. There are six frequency–response curves in the middle graph, which are visualized after local enlargement. It is evident that the curves are sparser after the sampled feature transform. The Fourier transform results are visualized as shown on the right of Figure 7a after the reconstruction of the feature map information from the attentional up-sampling. It can be observed that there is a richer frequency response in the middle- and high-frequency parts, while the logarithmic amplitude intensity is more concentrated. The result is consistent with the literature [42], in which it was verified that the multi-headed self-attentive biased low-pass filter behaves differently. Dynamic sampling is followed by upsampling, which compensates for the response of the high-frequency features. Comparing the global sparse sampling shown in row (b), where the left panel shows the Fourier variation results after global sparse sampling of the self-attention module, it can be seen that in the middle panel after local scaling, there are more frequency response curves than dynamic sparse sampling. Their different sampling methods lead to differences in the amplitude range of the feature values after Fourier transform, while the reconstructed transform graph on the right is consistent with row (a) in terms of frequency change trend and feature performance. The proposed method restores information with sparser sampling points and reduces the representation of redundant features, which could allow more information representation to be obtained with the same number of samples.

In this paper, dynamic sampling is used in the multi-head attention head to obtain representative feature points to sparse the query matrix to enhance the attention-modeling effect. The visualized sampled points are shown in Figure 8.

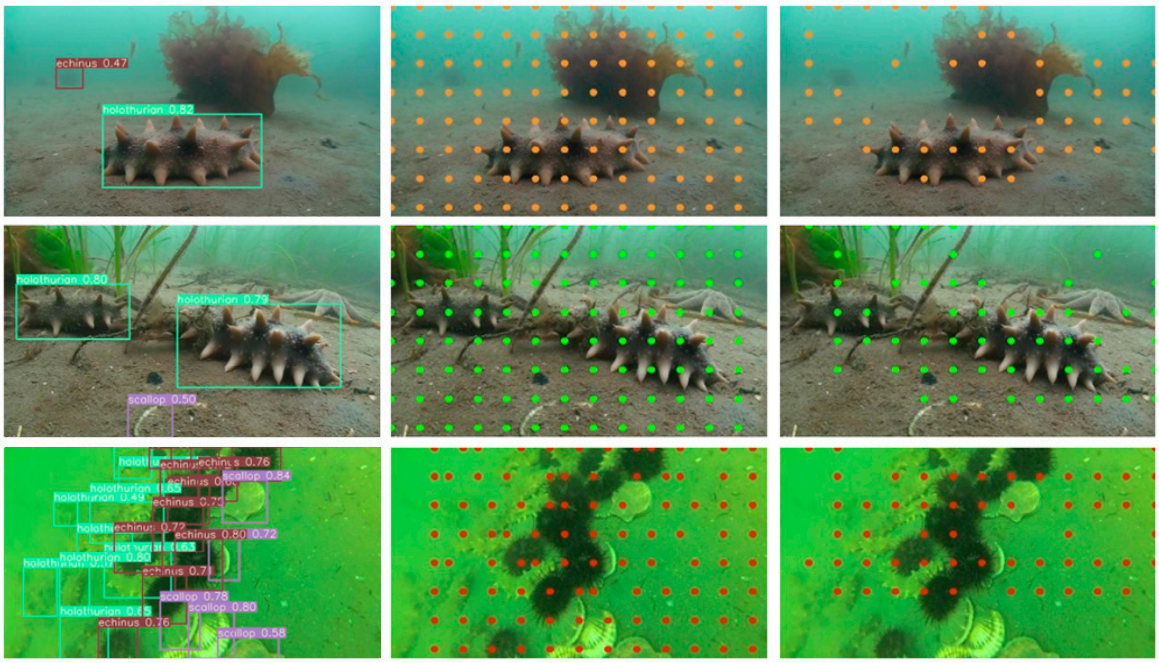

Figure 8. Dynamic sampling point visualization.

The left figure shows the visualization of the detection effect. The middle plot shows the self-attention sampling-point results from the middle layer's feature fusion module, representing the offset sampling results for medium-scale targets. As in row 3 in the dense target, the sampled points are shifted toward the target out, while uniform sampling is maintained in the background. The right side of the figure shows the self-attentive sampling results of the high-level feature fusion module, which represents the offset sampling results of the large-scale targets. For example, rows 1 and 2 have more densely sampled points on large targets.

The dynamic sampling points of underwater targets in different scenarios are sparser, and the sampling locations are more representative. There are more sampled points in the image target and representative background, which gives spatial semantics to the sampled feature points after the feature-aggregation module and improves the attention-modeling effect.

4.4. Ablation Experiments

Downsampling the feature map could enable the multi-headed attention to establish a sparsification matrix, which contributes to the performance improvement. Moreover, the dynamic sampling of Q enables the selection of focused aggregation regions. Computation with globally sampled K and V can enrich the communication between features. In this paper, we conduct ablation experiments on the sampling method and distillation method to verify the effectiveness of the dynamic sampling method.As shown in Table 2.

Table 2. Results of the ablation experiment with the reason-Transformer sampling method.

Method	mAP	Param (M)	Inference Memory (Batch = 4)	GFLOPs
Original Network	72.5	1.57	2138 MB	4.0
Reason-Transformer	74.3	2.14	4622 MB	4.5
Uniform sparse sampling	73.6	2.10	2510 MB	4.8
Dynamic sparse sampling	74.1	2.10	2522 MB	4.9

Building the Reason Layer with a standard Transformer in the original network could compensate for the lack of global attention in YOLOv5-lite. Compared to the original network, mAP has been improved by 1.8%, while the number of parameters grows by 589,984, and such overhead could be enhanced by adding read-only memory in the underwater device. However, the memory (RAM) consumed by its inference grows 2484 MB, and the underwater device computational unit has difficulty adapting to this resource consumption. The feature cache for forward inference is reduced by 45.69% by uniformly sparse sampling, while the mAP is improved by 1.1%. Applying the dynamic sparse sampling Reason Layer proposed in this paper can improve the mAP by 1.6% while maintaining the same number of parameters, which is more suitable for target detection in an underwater environment.

To verify the effectiveness of distillation losses in this paper, ablation experiments were conducted using different classification losses for the gridded localization distillation, the feature-map-distillation method, and the direct-response distillation.

As shown in Table 3 above, the direct response distillation method enhances the performance of the original network. However, the direct-response distillation method does not improve well due to the lack of reasonable localization knowledge. Feature map consistent distillation as the mainstream distillation strategy improves the student network by 1.8%, but this method requires caching the feature layers of the teacher network at three resolutions and regressing them with the corresponding features of the student network for consistency, which takes up more training resources. YOLOv5-lite takes the feature map centroid as the target-prediction Anchor point and undertakes the main localization task, so the centroid location is distilled separately, and its results are improved by 1.3%. Distillation was performed using Anchor boxes with grid-filter 8 neighborhoods and Anchor boxes with grid-filter 4 neighborhoods determined with target confidence, respectively, and we found that they obtained the same results. The mAP is also improved by 1.7% when the grid filtering method described in our paper is used to extract the teacher network's real labeling feature points and perform neighborhood grid distillation. The aforementioned comparison indicates that our method could use the surrounding valid prior boxes for auxiliary regression, remove the interference of invalid candidate boxes, and increase the knowledge of positive samples. Finally, the network achieves a performance similar to that of the feature-map consistency method with a lower training cache. It is worth mentioning that because the feature map consistency method makes the feature values used to predict branches consistent, using both methods makes it difficult to achieve a large performance improvement.

Table 3. Results of ablation experiments with different knowledge-distillation methods.

Center Point	Grid Filter 4 Neighborhoods	Grid Filter 8 Neighborhoods	Direct Response Distillation	Feature Map Distillation	mAP
-	-	-	-	-	72.5
✓	-	-	-	-	73.8
-	✓	-	-	-	74.2
-	-	✓	-	-	74.2
-	-	-	✓	-	73.5
-	-	-	-	✓	74.3
-	✓	-	-	✓	74.3

For the problem of the difficulty obtaining an equilibrium class of underwater data samples, the combination of different loss functions is applied to ablate the proposed method in our paper, and the results are summarized as follows.

It can be observed from Table 4 that it is difficult to adapt the originally adopted cross-entropy loss function to the unbalanced categories. By using the sample loss (focal) with optimized difficulty, the AP of category 1 improves by 1.5%, and the AP of category 4 improves by 0.7%. After applying polynomial loss (poly) to the network, the AP of category 1 improves by 1.5%, the AP of category 4 improves by 0.6%, and the AP of other categories

slightly decays. Applying the combination of the distillation strategies proposed in this paper resulted in a 6.6% AP improvement in category 1 and a 1% AP improvement in category 4, along with a slight AP improvement for the other categories.

Table 4. Comparison of experimental results of different recognition-loss functions.

Class-Loss	Class 1	Class 2	Class 3	Class 4
	AP			
CE	61.2	90.6	65.2	80.2
Focal	62.7	90.6	64.4	80.9
Ploy	63.5	88.7	64.2	80.8
Ploy-distill	67.8	90.9	66.9	81.2

To evaluate the training cost, a graph was plotted by recording the training versus mAP values, and it was used to analyze the variation in different distillation strategies and classification methods in training, as shown in Figure 9.

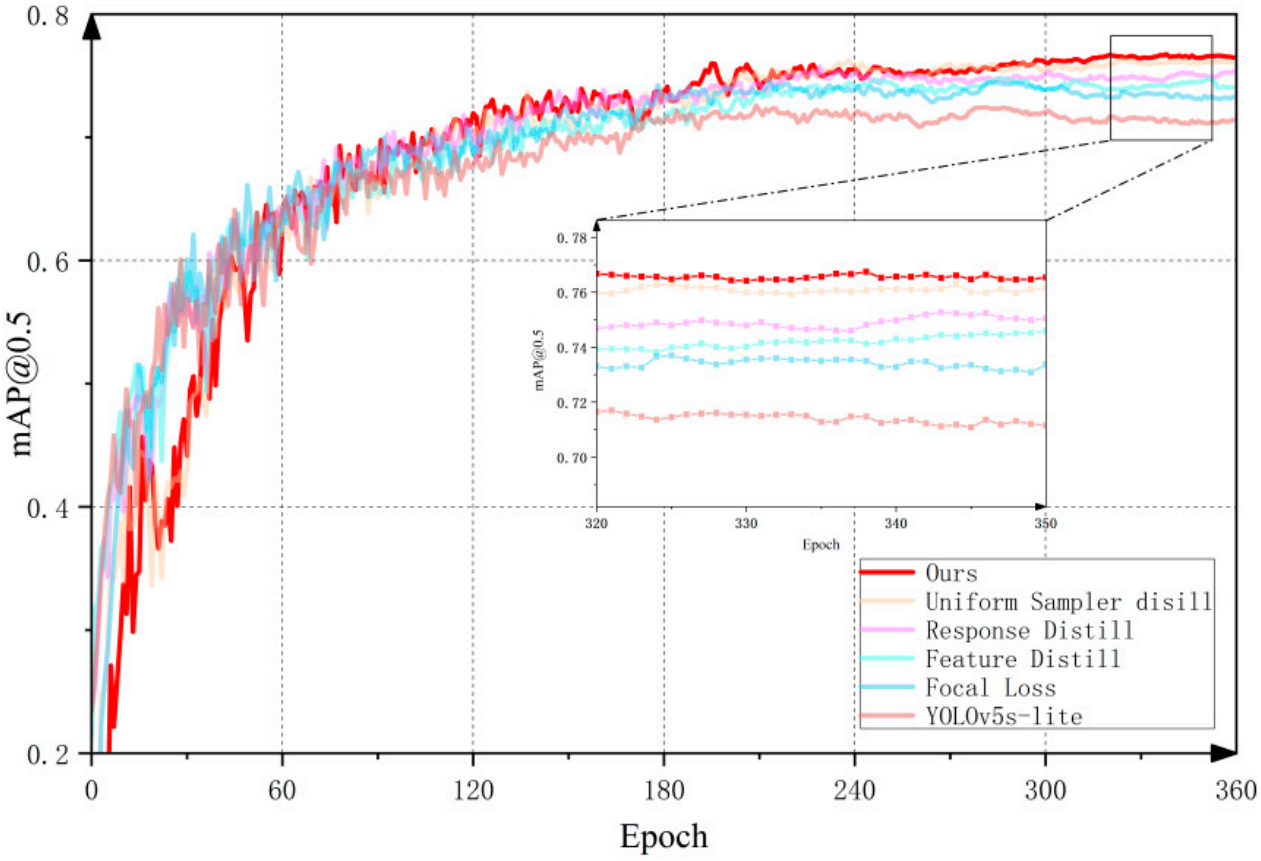

Figure 9. Training situation chart.

The models' performance metrics increased with the number of training iterations. After approximately 90 rounds, the performance advantages and disadvantages of various algorithmic strategies started to be gradually reflected. It can be observed that the proposed method could obtain optimal results after the final convergence by applying the grid localization distillation method to provide positive samples with different perspectives for localization knowledge to help the Anchor box with regression. The CE Loss and Focal Loss improve the performance to a limited extent. Due to the complexity of the underwater environment, there is a tendency for the mAP to decrease in the late training period. The method in this paper improves the performance relatively slowly in the early stage, during which a large amount of uncertain knowledge is accumulated, but the positive sample

regression gradually becomes accurate in the middle of training and rapidly improves the performance.

To verify the overall performance of our proposed algorithm, for the experiments, we selected the current mainstream deep-learning-based target-detection neural network as a reference, compared the mAP values detected by the algorithm through the same hardware platform, and tested them with the same data set. The experimental results are shown in Table 5.

Table 5. Mainstream algorithm comparison experimental results.

Method	Time Spent in Detection (ms)	Size	mAP	Parameters (M)
Faster RCNN (Vgg)	35.7	300 × 300	67.4	134.7
Faster RCNN (Resnet)	-	512 × 512	73.2	-
SA-FPN	-	1280 × 768	75.3	-
R-FCN	25.2	1000 × 600	75.2	31.9
MobileNet-SSD	-	300 × 300	61.2	5.50
SSD	21.2	300 × 300	64.5	24.2
RetinaNet	-	600 × 600	68.9	53.3
CenterNet	21.8	512 × 512	73.57	32.69
YOLOv4	22.8	416 × 416	78.0	61.38
YOLOv7	21.4	416 × 416	79.1	36.9
YOLOv7-tiny	21.2	416 × 416	77.9	6.2
YOLOv5s	20.7	416 × 416	76.7	6.74
YOLOv5l	21.7	416 × 416	77.2	44.40
YOLOv5s-lite	20.5	416 × 416	70.1	1.57
YOLOv5s-lite (ours)	20.7	416 × 416	76.7	2.10
YOLOv5g-lite	20.6	416 × 416	73.8	5.32
YOLOv5g-lite (ours)	21.0	416 × 416	78.8	5.68

The proposed algorithm showed relatively high accuracy in single-stage algorithms, with 12.2% and 15.5% improvement compared to SSD and its lightweight improvements. The detection accuracy was improved by 3.14% compared to the Anchor-free detection network CenterNet and 3.5% compared to the two-stage network Faster RCNN. Among the related improvements to the YOLO series, the algorithm in this paper can improve in detection speed over YOLOv4 by 75% and a detection speed 58.8% faster than YOLOv5-l. It also has an advantage over the tiny model of the YOLOv7 algorithm [37] in terms of accuracy and number of parameters. It improves accuracy over the lite type large volume network YOLOv5g-lite by 5%, achieves the same accuracy as the standard YOLOv5s with the lite-type minimal model, and improves over the baseline network mAP by 6.6%.

5. Discussion

The perception of underwater robots guides the subsequent actions, and their perception models require light weight and high localization accuracy. The single-stage detection algorithm ignores the importance of global information when constructing feature fusion modules. The ViT (Vision Transform, ViT) detection algorithm has received more attention for its ability to construct global information associations, but it is unsuitable for underwater device deployment due to its large number of parameters. In this paper, we propose a novel detection model with YOLOv5-lite combined with a Transformer and verify the effectiveness of strengthening the global information at the deeper level of the model through experiments and feature analysis. The proposed algorithm maintains the lightweight model and improves the detection accuracy, and the effective mechanism is considered in this paper to achieve dynamic sampling of the attention matrix in multi-headed attention. Adaptive feature encoding of important local blocks in deep features could be performed, which reduces the cost of matrix construction while acquiring important local features. The connection between deep features can be effectively fused in Neck, eventually improving the detection accuracy. In addition, this paper proposes a distillation method with decoupled localization and recognition. The algorithm could effectively handle regression

and recognition information by filtering in the grid. The output response of the student model was trained by distillation using the uncertainty of the confidence prediction. In this paper, we discovered that the knowledge information passed separately improves the detection ability of YOLO's prediction head, while the use of uncertainty allowed the accumulation of effective knowledge early in the training, ensuring the model eventually improves performance without increasing complexity.

6. Conclusions

Current application scenarios of deep learning are limited by the devices' computing power. The perceptual processing units of underwater robots cannot provide complex model-computing capabilities. Improving the models' detection accuracy while keeping the low complexity of the algorithms is a challenging problem. The key to underwater optical image detection is to equip the model with the ability to adapt to complex underwater environments. In this paper, we study the lite-type algorithm in YOLOv5. A dynamic sampling Transformer module is proposed deep in the model, which optimizes the coding process and saves forward propagation memory overhead. The sparsification of the feature matrix improves the perception of the features at the region of interest at the same time, which makes the model more easily detect targets in complex environments. Thereafter, this paper proposes a distillation strategy for decoupling localization and recognition knowledge in the multi-scale feature map using feature-map grids. It give the model more positive sample target position information for model training in underwater environments and improves the algorithm's target-localization accuracy. Meanwhile, the gradient's higher-order term is introduced in the loss function to improve the accuracy of the model in unbalanced detection of underwater data categories. Finally, the proposed algorithm maintains a low number of parameters to improve the detection efficiency and detection accuracy in the underwater environment.

In future work, more data-expanded training strategies and post-processing methods will be investigated. The stability of the model under different noisy data should be examined as a way to compensate the problem of difficult access to underwater data.

Author Contributions: Conceptualization, L.C. and Y.Y.; Methodology, L.C.; Software, Y.Y.; Validation, Y.Y. and L.C.; Formal Analysis, J.Z.; Investigation, J.Z.; Resources, J.Z.; Data Curation, Z.W.; Writing—Original Draft Preparation, Y.Y.; Writing—Review and Editing, L.C. and Z.W.; Visualization, Y.Y.; Supervision, S.Z. and L.W.; Project Administration, L.W.; Funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62271199 and Science and Technology Talents Sponsorship Program by Hunan Association for Science and Technology, China, grant number 2022TJ-Q03.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Zhou, T.; Si, J.; Wang, L.; Xu, C.; Yu, X. Automatic Detection of Underwater Small Targets Using Forward-Looking Sonar Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4207912. [CrossRef]
2. Wan, Z.; Zhang, L.; Huang, H.; Yang, X. GSDCN: A Customized Two-Stage Neural Network for Benthonic Organism Detection. In *The Neural Information Processing*; Springer: Cham, Switzerland, 2020; pp. 811–820.
3. Zhou, J.; Yao, J.; Zhang, W.; Zhang, D. Multi-scale retinex-based adaptive gray-scale transformation method for underwater image enhancement. *Multimedia Tools Appl.* **2022**, *81*, 1811–1831. [CrossRef]
4. Liu, P.; Hongbo, Y.; Hu, Y.; Fu, J. Research on target recognition of underwater robot. In Proceedings of the 2018 IEEE International Conference on Advanced Manufacturing (ICAM), Yunlin, Taiwan, 16–18 November 2018; pp. 463–466.
5. Sarkar, P.; De, S.; Gurung, S. A Survey on Underwater Object Detection. In *Intelligence Enabled Research: DoSIER*; Springer: Singapore, 2021; pp. 91–104.

6. Chen, L.; Zhou, F.; Wang, S.; Dong, J.; Li, N.; Ma, H.; Wang, X.; Zhou, H. SWIPENET: Object detection in noisy underwater scenes. *Pattern Recognit.* **2022**, *132*, 108926. [CrossRef]
7. Zhang, X.; Fang, X.; Pan, M.; Yuan, L.; Zhang, Y.; Yuan, M.; Lv, S.; Yu, H. A Marine Organism Detection Framework Based on the Joint Optimization of Image Enhancement and Object Detection. *Sensors* **2021**, *21*, 7205. [CrossRef] [PubMed]
8. Wang, J.; He, X.; Shao, F.; Lu, G.; Jiang, Q.; Hu, R.; Li, J. A Novel Attention-Based Lightweight Network for Multiscale Object Detection in Underwater Images. *J. Sens.* **2022**, *2022*, 2582687. [CrossRef]
9. Feng, H.; Xu, L.; Yin, X.; Chen, Z. Underwater salient object detection based on red channel correction. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 26–28 March 2021; pp. 446–449.
10. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the 30th International Conference on Neural, Barcelona, Spain, 5–10 December 2016. [CrossRef]
11. Xu, F.; Wang, H.; Peng, J.; Fu, X. Scale-aware feature pyramid architecture for marine object detection. *Neural Comput. Appl.* **2021**, *33*, 3637–3653. [CrossRef]
12. 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.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
13. 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, 11–14 October 2016; Volume 9905, pp. 21–37. [CrossRef]
14. 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]
15. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577. [CrossRef]
16. Aksoy, T.; Halici, U. Analysis of visual reasoning on one-stage object detection. *arXiv* **2022**, arXiv:2202.13115. [CrossRef]
17. Zhu, H.; Xie, Y.; Huang, H.; Jing, C.; Rong, Y.; Wang, C. DB-YOLO: A Duplicate Bilateral YOLO Network for Multi-Scale Ship Detection in SAR Images. *Sensors* **2021**, *21*, 8146. [CrossRef] [PubMed]
18. Lin, W.H.; Zhong, J.X.; Liu, S.; Li, T.; Li, G. ROIMIX: Proposal-Fusion Among Multiple Images for Underwater Object Detection. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2588–2592.
19. Jiang, L.; Wang, Y.; Jia, Q.; Xu, S.; Liu, Y.; Fan, X.; Wang, R. Underwater Species Detection using Channel Sharpening Attention. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 4259–4267.
20. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *Multimed. Tools Appl.* **2021**, *80*, 4037–4051.
21. Chen, G.; Choi, W.; Yu, X.; Han, T.; Chandraker, M. Learning efficient object detection models with knowledge distillation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 742–751.
22. Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Wang, Z.; Liu, S.; Zhou, E. General instance distillation for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021. [CrossRef]
23. Pang, J.; Liu, W.; Liu, B.; Tao, D.; Zhang, K.; Lu, X. Interference Distillation for Underwater Fish Recognition. In Proceedings of the 6th Asian Conference on Pattern Recognition, Jeju, Republic of Korea, 9–12 November 2021; pp. 62–74.
24. Zheng, Z.; Ye, R.; Hou, Q.; Ren, D.; Wang, P.; Zuo, W.; Cheng, M.-M. Localization distillation for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9397–9406.
25. Ji, W.; Peng, J.; Xu, B.; Zhang, T. Real-time detection of underwater river crab based on multi-scale pyramid fusion image enhancement and MobileCenterNet model. *Comput. Electron. Agric.* **2023**, *204*, 107522. [CrossRef]
26. Han, Y.; Chen, L.; Luo, Y.; Ai, H.; Hong, Z.; Ma, Z.; Wang, J.; Zhou, R.; Zhang, Y. Underwater Holothurian Target-Detection Algorithm Based on Improved CenterNet and Scene Feature Fusion. *Sensors* **2022**, *22*, 7204. [CrossRef] [PubMed]
27. Huang, A.; Zhong, G.; Li, H.; Choi, D. Underwater Object Detection Using Restructured SSD. In Proceedings of the CAAI International Conference on Artificial Intelligence, Beijing, China, 27–28 August 2022; pp. 526–537.
28. Dinakaran, R.; Zhang, L.; Li, C.-T.; Bouridane, A.; Jiang, R. Robust and Fair Undersea Target Detection with Automated Underwater Vehicles for Biodiversity Data Collection. *Remote Sens.* **2022**, *14*, 3680. [CrossRef]
29. Wang, X.; Lin, J.; Zhao, J.; Yang, X.; Yan, J. EAutoDet: Efficient Architecture Search for Object Detection. In *European Conference on Computer Vision, Glasgow, UK, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 668–684.
30. Bancud, G.E.; Labanon, A.J.; Abreo, N.A.; Kobayashi, V. Combining Image Enhancement Techniques and Deep Learning for Shallow Water Benthic Marine Litter Detection. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Turin, Italy, 18–22 September 2023*; Springer Nature: Cham, Switzerland, 2023; pp. 137–149. [CrossRef]
31. Chen, L.; Yang, Y.; Wang, Z.; Zhang, J.; Zhou, S.; Wu, L. Underwater Target Detection Lightweight Algorithm Based on Multi-Scale Feature Fusion. *J. Mar. Sci. Eng.* **2023**, *11*, 320. [CrossRef]
32. Liu, Z.; Zhuang, Y.; Jia, P.; Wu, C.; Xu, H.; Liu, Z. A Novel Underwater Image Enhancement Algorithm and an Improved Underwater Biological Detection Pipeline. *J. Mar. Sci. Eng.* **2022**, *10*, 1204. [CrossRef]
33. Zhao, S.; Zheng, J.; Sun, S.; Zhang, L. An Improved YOLO Algorithm for Fast and Accurate Underwater Object Detection. *Symmetry* **2022**, *14*, 1669. [CrossRef]

34. Guo, J.; Han, K.; Wang, Y.; Wu, H.; Chen, X.; Xu, C.; Xu, C. Distilling object detectors via decoupled features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021. [CrossRef]
35. Tang, S.; Zhang, Z.; Cheng, Z.; Lu, J.; Xu, Y.; Niu, Y.; He, F. Distilling Object Detectors with Global Knowledge. In *European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 422–438.
36. Yang, C.; Ochal, M.; Storkey, A.J.; Crowley, E.J. Prediction-Guided Distillation for Dense Object Detection. In *European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 123–138.
37. 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. [CrossRef]
38. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European conference on computer vision (ECCV), Montreal, QC, Canada, 11 October 2021; pp. 122–138.
39. Pan, J.; Bulat, A.; Tan, F.; Zhu, X.; Dudziak, L.; Li, H.; Martinez, B. EdgeViTs: Competing Light-weight CNNs on Mobile Devices with Vision Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 294–311.
40. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
41. Leng, Z.; Tan, M.; Liu, C.; Cubuk, E.D.; Shi, X.; Cheng, S.; Anguelov, D. PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022. [CrossRef]
42. Park, N.; Kim, S. How Do Vision Transformers Work? In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022. [CrossRef]