Journal of
Marine Science and Engineering

MDPI

Article

MTP-YOLO: You Only Look Once Based Maritime Tiny Person Detector for Emergency Rescue

Yonggang Shi [1], Shaokun Li [1], Ziyan Liu [1], Zhiguo Zhou [1,2,*] and Xuehua Zhou [1,2,*]

[1] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China; ygshi@bit.edu.cn (Y.S.); 3120221325@bit.edu.cn (S.L.); 1120200666@bit.edu.cn (Z.L.)

[2] Tangshan Research Institute, Beijing Institute of Technology, Tangshan 063000, China

* Correspondence: zhiguozhou@bit.edu.cn (Z.Z.); xuehuazhou@bit.edu.cn (X.Z.)

Abstract: Tiny person detection based on computer vision technology is critical for maritime emergency rescue. However, humans appear very small on the vast sea surface, and this poses a huge challenge in identifying them. In this study, a single-stage tiny person detector, namely the "You only look once"-based Maritime Tiny Person detector (MTP-YOLO), is proposed for detecting maritime tiny persons. Specifically, we designed the cross-stage partial layer with two convolutions Efficient Layer Aggregation Networks (C2fELAN) by drawing on the Generalized Efficient Layer Aggregation Networks (GELAN) of the latest YOLOv9, which preserves the key features of a tiny person during the calculations. Meanwhile, in order to accurately detect tiny persons in complex backgrounds, we adopted a Multi-level Cascaded Enhanced Convolutional Block Attention Module (MCE-CBAM) to make the network attach importance to the area where the object is located. Finally, by analyzing the sensitivity of tiny objects to position and scale deviation, we proposed a new object position regression cost function called Weighted Efficient Intersection over Union (W-EIoU) Loss. We verified our proposed MTP-YOLO on the TinyPersonv2 dataset. All these results confirm that this method significantly improves model performance while maintaining a low number of parameters and can therefore be applied to maritime emergency rescue missions.

Keywords: tiny person detection; cross-stage partial layer with two convolutions efficient layer aggregation networks; multi-level cascaded enhanced convolutional block attention module; weighted efficient intersection over union

check for updates

Citation: Shi, Y.; Li, S.; Liu, Z.; Zhou, Z.; Zhou, X. MTP-YOLO: You Only Look Once Based Maritime Tiny Person Detector for Emergency Rescue. *J. Mar. Sci. Eng.* **2024**, *12*, 669. https://doi.org/10.3390/jmse12040669

Academic Editor: Gerasimos Theotokatos

Received: 19 March 2024
Revised: 10 April 2024
Accepted: 10 April 2024
Published: 18 April 2024

1. Introduction

With increasing global maritime activities, the complexity and urgency of maritime rescue missions have also increased. In this context, quickly and accurately locating people in distress is the key to improving rescue efficiency and minimizing casualties and property losses. However, existing methods such as manual observation or satellite positioning still face many challenges in accurately locating victims. Manual observation can easily cause personnel fatigue and distraction, thereby increasing the risk of missing search and rescue targets. For satellite positioning, when the signal quality is poor or the victim's communication equipment fails, positioning cannot be performed. Therefore, existing detection technology cannot meet the requirements of modern maritime emergency rescue.

Lately, there has been a growing emphasis among scholars on techniques for object detection through visual data analysis. As it uses high-resolution cameras and complex image processing algorithms to identify objects in specific scenes, this method will not be affected by visual fatigue and signal quality when being used for maritime emergency rescue. Vision-based object detection algorithms mainly include the following categories: two-stage object detectors represented by the RCNN series, one-stage detectors represented by the YOLO series [1], and Transformer-based DETR series [2]. Although they all achieved impressive performances with natural images, detecting tiny objects at sea remains a challenge [3].

On the one hand, there is a certain commonality between maritime tiny object detection and small object detection. Firstly, compared to regular images, tiny objects in sea surface images are smaller in size and contain less information compared to the entire image. Secondly, limited information about tiny objects may disappear during forward propagation, and the model may not be able to capture the key features of tiny objects, leading to detection errors. Furthermore, tiny objects might overlap with each other, challenging the detector's ability to differentiate between objects that are close together.

On the other hand, detecting small objects at sea has its own unique characteristics. For example, the lighting conditions at sea may be complex and variable due to the presence of more specular reflections [4]. The fluctuation in lighting intensity can significantly impact the imaging effect of the camera. This complicates the process of discerning the characteristics of the object. Moreover, due to the lack of additional light sources, the lighting conditions largely depend on sunlight, resulting in a large number of backlit scenes. In scenes with such backlighting, there is a scenario where the object is poorly lit, contrasted by a very bright background, and both factors might potentially degrade the efficacy of object detection.

To tackle these particular hurdles, we designed an innovative architecture named MTP-YOLO for tiny person detection in maritime emergency rescue missions. We trained and evaluated MTP-YOLO on the TinyPersonv2 [5] dataset, which contains sea surface images annotated with tiny person labels. The results indicate that MTP-YOLO can improve the detection ability of tiny objects compared to the most advanced methods currently available. Our contributions are summarized as follows:

1. We designed a new feature extraction module called C2fELAN to better retain tiny object information and reduce information loss during forward propagation, allowing the model to use this information to detect tiny objects and overcome the challenges of tiny object detection.
2. We adopted the Multi-level Cascaded Enhanced CBAM to obtain a more focused attention distribution, allowing the model to attach importance to areas where the important features of tiny objects exist and learn more useful information.
3. We proposed a new bounding box regression loss function called Weighted EIoU Loss to solve the problem of tiny objects having different sensitivities to position and scale deviation and boost the model's performance in identifying tiny persons.

2. Related Work

2.1. Object Detection

At present, the popular object detection algorithms mainly include the following categories: two-stage object detectors represented by the RCNN series, one-stage detectors represented by the YOLO series [6], and DETR series based on the Transformer. After integrating the RPN [7] structure, the RCNN series algorithm greatly improves detection accuracy but is slow and cannot fulfill the demands of real-time detection in most applications. Algorithms in the YOLO series approach the task of object detection as a problem of spatial regression. It uses CSPNet [8], PAN [9], FPN [10], and a series of their variants as the basic building blocks of the network. While fulfilling real-time detection requirements, its accuracy reaches the same level as the RCNN series. The DETR series of algorithms are introduced, utilizing the Transformer architecture [11] from the domain of natural language processing; however, it is difficult to be applied to new fields without a pre-trained model in the corresponding field. Therefore, the YOLO series currently remains the most widely used algorithm. YOLOv8 [12] was chosen as the basis of this article as this method has proven to be powerful in a large number of computer vision tasks.

2.2. Tiny Object Detection

Despite significant advancements in the field of object detection algorithms, research on tiny object detection still faces great challenges, including the following: (1) the tiny object itself occupies a small size in the image and has limited available information;

(2) the features of tiny objects may disappear during the forward propagation process of the network, which poses certain difficulties to detection; (3) tiny objects in complex environments will be interfered with by factors like lighting, occlusion, and aggregation, thereby complicating their differentiation from the backdrop or akin items. In response to the difficulties in tiny object detection, researchers have made numerous improvements to mainstream object detection algorithms. Their improvement methods can be divided as follows: context information learning methods [13] that solve the problem of limited feature information being carried by tiny objects, multi-scale feature fusion methods [14] that integrate multiple feature layers to improve the representation ability of tiny objects, and attention mechanism methods to [15–18] improve the model's attention to tiny object features. Although these works have improved tiny object detection performance in their respective scenarios, they may not be applicable when there are scene changes. Considering this, we propose that the MTP-YOLO algorithm is suitable for the scenarios listed in this paper.

3. Method

3.1. Overview of MTP-YOLO

MTP-YOLO is built upon YOLOv8 and comprises three primary parts, as depicted in Figure 1. The backbone of MTP-YOLO includes the following four main components: CBS (conv2d, batch normalization, sigmoid linear unit), C2fELAN, MCE-CBAM, and SPPF (Spatial Pyramid Pooling Fast). These are utilized to extract pertinent attributes of the object from an input image. The neck architecture still adopts the FPN (Feature Pyramid Network) and PAN (Path Aggregation Network) structures for feature fusion at different scales, mainly composed of C2fELAN, Concat, Upsample, and CBS components. This enhances the model's ability to recognize objects of various sizes by integrating localization information and semantic information. The head module adopts the current mainstream decoupling head structure, dividing the detection head into a regression branch and a classification branch. The regression branch uses DFL (Distribution Focal Loss) and Weighted EIoU Loss, while the classification branch uses BCE (Binary Cross Entropy) loss.

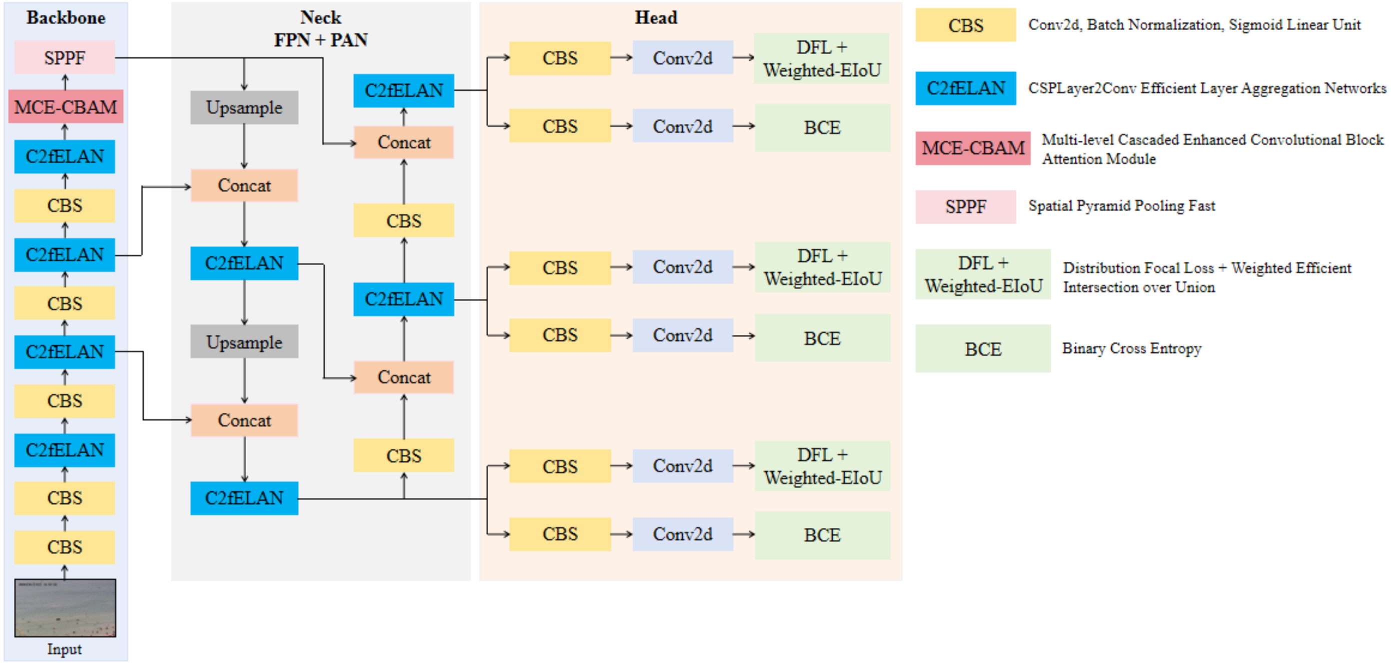

Backbone
SPPF
MCE-CBAM
C2fELAN
CBS
C2fELAN
CBS
C2fELAN
CBS
C2fELAN
CBS
CBS
Input
Neck
FPN + PAN
Upsample
Concat
C2fELAN
Upsample
Concat
C2fELAN
C2fELAN
Concat
CBS
C2fELAN
Concat
CBS
Head
CBS
Conv2d
DFL + Weighted-EIoU
BCE
CBS
Conv2d, Batch Normalization, Sigmoid Linear Unit
C2fELAN
CSPLayer2Conv Efficient Layer Aggregation Networks
MCE-CBAM
Multi-level Cascaded Enhanced Convolutional Block Attention Module
SPPF
Spatial Pyramid Pooling Fast
DFL + Weighted-EIoU
Distribution Focal Loss + Weighted Efficient Intersection over Union
BCE
Binary Cross Entropy

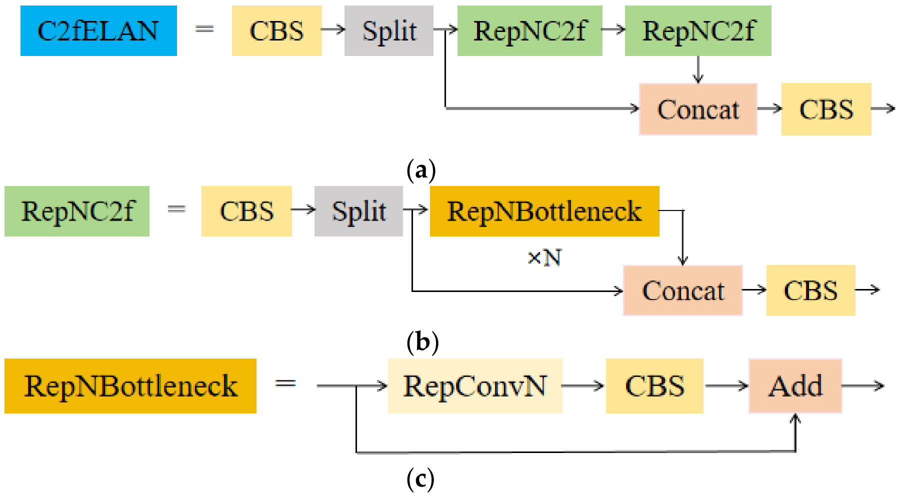
Figure 1. Overview of MTP-YOLO.

MTP-YOLO retains the overall style of YOLOv8. However, the original YOLOv8 network was not designed for tiny object detection tasks, which reduces its applicability in tiny person detection tasks at sea. Therefore, we improved the network structure to enhance its performance in detecting tiny objects. Firstly, we drew inspiration from the GELAN structure of the latest object detector, YOLOv9 [19], and design the C2fELAN module to replace the C2f modules in the original YOLOv8 backbone and neck. This allowed us to preserve the key features of the tiny objects and obtain richer gradient information during the calculation process, thereby achieving a higher detection accuracy. Secondly, we inserted our proposed MCE-CBAM module before the SPPF layer in the YOLOv8 architecture to boost the feature extraction ability of the backbone and make the model pay attention to key details that are conducive to identifying tiny objects, thereby improving the ability to detect tiny objects. In addition, by analyzing the sensitivity of tiny objects to position and scale deviation, we designed a new boundary box regression loss called Weighted EIoU Loss, substituting the CIoU in the cost function to alleviate the substantial influence of tiny object position deviation on detection performance, thereby improving the detection performance of tiny objects.

3.2. C2fELAN Module

By combining two neural network architectures designed using gradient path planning, CSPNet (Cross Stage Partial Network) and ELAN [20], the authors of YOLOv9 designed a Generalized Efficient Layer Aggregation Network (GELAN) that considers weight, inference speed, and accuracy. The design purpose of CSPNet is to enable the network to obtain richer gradient fusion information while reducing computational complexity. The method divides the tensor of the base layer into two segments, which are then merged through a cross-stage hierarchical approach. By separating the gradient flows, they can propagate on different network paths. In addition, CSPNet can greatly reduce computational complexity, and improve inference speed and accuracy. The main purpose of designing ELAN is to address the issue of the gradually deteriorating convergence of deep models during model scaling. Comparing VoVNet (Variety of View Network) and ResNet (Residual Network), VoVNet performs worse than ResNet when stacking more blocks. The authors analyzed that this is because there are too many transition layers in the VoVNet structure, which leads to an increasing number of shortest gradient paths when stacking blocks, making training more difficult as the number of blocks increases. Therefore, by appropriately deleting the transition layer, network performance can be improved, and the shortest gradient path of the entire network can be quickly lengthened. When the network is stacked deeper, the above design strategy can then successfully train ELAN. The author of YOLOv9 extended the ability of ELAN, which initially only used convolutional layers for stacking, to a new architecture that can accommodate any kind of computational block.

Taking inspiration from the GELAN module proposed by the author of YOLOv9, we combined the C2f and ELAN neural network modules with gradient path planning to design the feature extraction module, C2fELAN, used in this paper. This structure can retain relatively complete feature information of small objects and provide reliable gradient information that can be used to determine the objective function. The comprehensive layout is depicted in Figure 2. Specifically, we replaced the stacking of convolution modules in ELAN modules with the stacking of RepNC2f (re-parameterization cross stage partial layer with two convolutions without identity connection) modules. RepNC2f modifies the convolution in the bottleneck structure of the C2f module to RepConvN (re-parameterization convolution without identity connection), which is the structure of RepConv (re-parameterization convolution) after removing the identity mapping. The RepConv idea is to reparameterize the RepVGG block used during training, converting the 1×1 convolution and unprocessed identity maps in RepVGG into a 3×3 convolution, and then fusing them. By applying this RepConv to RepNC2f and RepNBottleneck, the inference efficiency of the network can be greatly improved.



(**a**)

(**b**)

(**c**)

Figure 2. The structure of C2fELAN and its components. (**a**) The details of C2fELAN; (**b**) The details of RepNC2f; (**c**) The details of RepNBottleneck. RepNC2f: re-parameterization cross stage partial layer with two convolutions without identity connection. RepNBottleneck: re-parameterization bottleneck without identity connection. RepConvN: re-parameterization convolution without identity connection.

As shown in Figure 3, when the model is in the training phase, the RepConvN module has two different convolution kernels: 3 × 3 and 1 × 1. When the model is in the inference stage, the 1 × 1 and 3 × 3 convolution kernels can be combined into a single 3 × 3 kernel through structural reparameterization. The specific method includes filling the surrounding parts of the 1 × 1 kernel into a 3 × 3 form. Based on the additivity principle of convolution kernels of the same size, the padding kernel is added to the original 3 × 3 convolution kernel to form a 3 × 3 convolution kernel for inference.

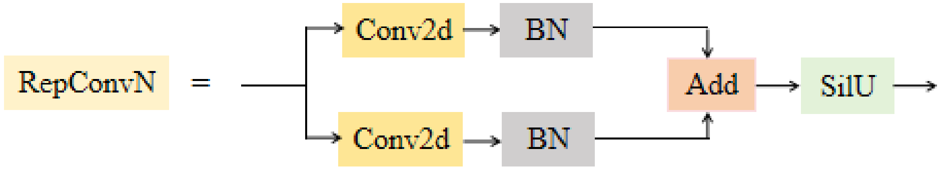


Figure 3. The details of RepConvN. BN: batch normalization. SilU: sigmoid linear unit.

3.3. Multi-Level Cascaded Enhanced CBAM Module

In recent years, various object detection architectures have adopted attention mechanisms to optimize their models and have achieved good results. Research on the combination of deep learning and visual attention mechanisms mostly focuses on using masks to form attention mechanisms. The principle of masking is to identify key features in image data through another layer of new weights. Through learning and training, deep neural networks learn the areas that need attention in each new image, forming the necessary attention. Among them, the most typical attention mechanisms include the self-attention, spatial attention, and temporal attention mechanisms. These attention mechanisms allow the model to assign different weights to different positions of the input sequence in order to focus on the most relevant part when processing each sequence element.

Therefore, MTP-YOLO within this article was also designed with a Multi-level Cascaded Enhanced CBAM, targeted at improving the tiny person detection effect, as depicted in Figure 4. Considering that the original CBAM can enhance the model's ability to focus on key features, we stacked and cascaded the spatial attention module and channel attention module in the CBAM to further enhance the model's performance to focus on crucial attributes and improve its detection performance.

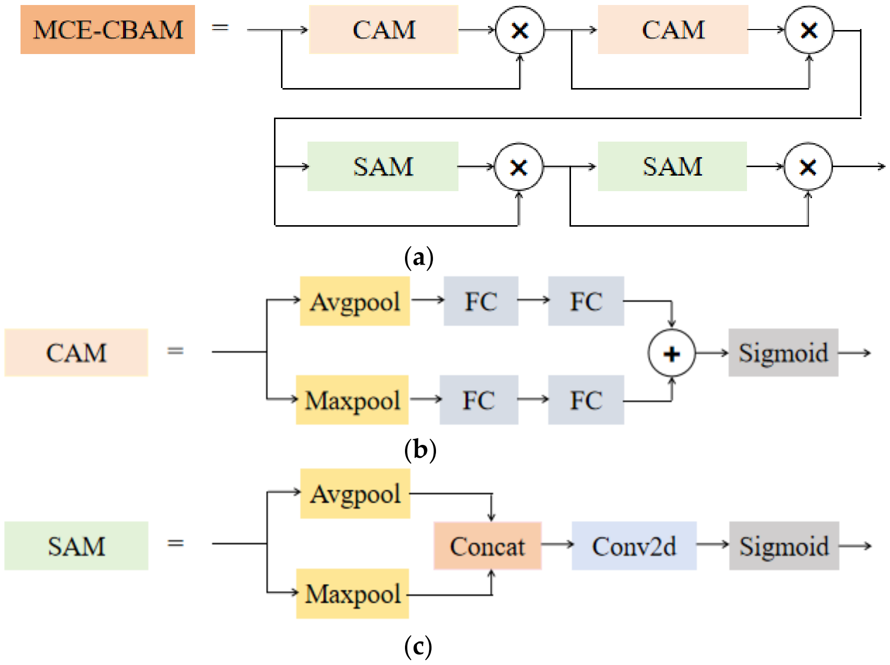


Figure 4. The structure of MCE-CBAM and its components. (**a**) The structure of MCE-CBAM; (**b**) The details of channel attention mechanism; (**c**) The details of spatial attention mechanism. CAM: channel attention module. SAM: spatial attention module. FC: fully connected layer.

The traditional CBAM aims to enhance the network's representation ability by introducing attention mechanisms, including the following two submodules: channel attention module and spatial attention module. Through the adaptive refinement of intermediary feature representations in each convolutional block of the deep network, CBAM achieves the attention to key information and suppression of unnecessary information. Utilizing the operations of both average and max pooling, the channel attention mechanism integrates the spatial information from the input feature maps, resulting in the acquisition of dual feature maps. After feeding them separately into a shared multi-layer perceptron, the output features of the two multi-layer perceptrons are added element by element, and the channel attention map is generated through a sigmoid activation function. The spatial attention mechanism first conducts channel-wise global maximum pooling and global average pooling on the input feature map, yielding a pair of feature maps. Next, these two feature maps along the channel axis are concatenated and a convolution is executed to reduce the parameter count. Subsequently, spatial attention features are generated through sigmoid operations.

3.4. Weighted-EIoU Loss

The loss associated with object position regression is a vital part of the loss function used in object detection, and currently the mainstream bounding box regression loss is the IoU series [21–24]. Although it has undergone multiple evolutions, we found that they have all overlooked a problem, which is the different sensitivities of tiny objects to the positional and scale deviations of the detection box, as shown in Figure 5. When the detection box is offset by a width in the horizontal direction, the tiny object will disappear from the detection box, resulting in a missed detection, even though the offset distance is very small. When the size of the box doubles, the tiny object is still in the detection box, and the model can still recognize the small object without missing the detection.

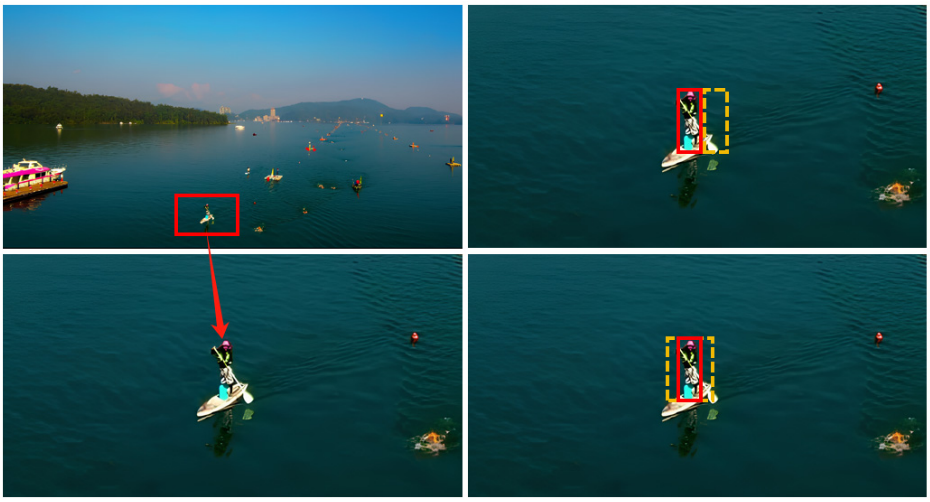


Figure 5. Comparison of the sensitivity of small objects to the position and scale deviation of bounding boxes. The red box represents the ground truth, and the yellow dashed box represents the detection box.

To address this problem, we propose a novel bounding box loss function called Weighted-EIoU Loss. We apply different weights to the center point distance deviation term and the detection box scale deviation term respectively, with the aim of boosting the detection efficacy of tiny objects:

$$WEIoU = IoU - \alpha\frac{\rho^2\left(b, b^{gt}\right)}{c^2} - \beta\frac{\rho^2\left(w, w^{gt}\right)}{C_w^2} - \beta\frac{\rho^2\left(h, h^{gt}\right)}{C_h^2} \tag{1}$$

$$\alpha + 2\beta = 3 \tag{2}$$

where IoU represents the intersection and union ratio of the detection box and the ground truth, c is the diagonal length of the minimum bounding rectangle, C_w and C_h are the width and height of the minimum bounding rectangle, b and b^{gt} represent the center points of the detection box and the ground truth, w and h represent the width and height of the detection box, w^{gt} and h^{gt} represent the width and height of the ground truth, ρ stands for the Euclidean distance, and α and β represent the weight applied. Equation (1) takes into account the overlapping area, center point distance, and differences in width and height between the predicted and actual boxes simultaneously. Among them, $\alpha + 2\beta = 3$ is meant to adjust only the weight within the bounding box loss, thereby avoiding implicitly imposing additional weights between the bounding box loss and the classification loss. In addition, considering that tiny objects are more sensitive to center point distance offset, we set $\alpha > \beta$ in the experiment.

4. Experiment

4.1. Datasets and Experimental Settings

The TinyPersonv2 dataset used in this paper includes 6278 images, which are taken from Internet platforms such as Baidu, YouTube, and Bing, as well as from cameras, and are specially designed for tiny person detection. We randomly divide it into a training set and a validation set with a ratio of 8:2. In order to further enrich the dataset and enhance the generalization ability of the model, we also utilize multiple data augmentation methods, including HSV transformation, shifting, and mosaic augmentation, etc.

We trained our MTP-YOLO on NVIDIA RTX3060 GPU and used the PyTorch 2.0.1 framework. All the networks we mentioned did not use pretrained weights and were trained from scratch. The total training time is 500 epochs. The starting learning rate was configured at 0.01, with a momentum of 0.937 and a weight decay coefficient of 0.0005. Like YOLOv8, we turned mosaic augmentation off during the final 10 epochs. This model uses the SGD optimizer. We then configured the input image dimensions to 640 by 640 pixels and established the batch size as 8.

4.2. Comparison of Different Weight Values for Weighted EIoU

In order to determine how to set the weights of the Weighted EIoU to achieve the optimal model performance, we started from 1 and continuously increased the value of α with a step size of 0.5. We trained the model under different α values and evaluated its performance. Table 1 shows the experimental results, which show that when $\alpha = 1$, W-EIoU degenerates into EIoU. When $\alpha > 1$ and $\alpha < 3$, the model performance improves to varying degrees, and the best performance is achieved when $\alpha = 3.0$. This is because as it approaches 3.0, the width and height related terms in W-EIoU are suppressed, increasing its attention to positional deviation, and allowing for more tiny objects to be detected. Therefore, we set α to 3.0, where β is 0.0. This indicated a substantial enhancement in the model's detection capabilities upon the elimination of terms associated with width and height, confirming that tiny objects have a high sensitivity to positional discrepancies. Specifically, when α = 3, the terms related to width and height are suppressed, and the mAP score reaches its maximum value. This means that for small targets, the position of the center point should be considered as the main factor, while width and height become irrelevant. At this point, Weighted EIoU is somewhat similar to DIoU, but unlike DIoU, α = 3 weight is applied to the terms related to the distance from the center point.

Table 1. The results when taking different values of α.

α	Precision	Recall	mAP@0.5
1.0	0.767	0.579	0.675
1.5	0.750	0.573	0.665
2.0	0.769	0.580	0.680
2.5	0.773	0.590	0.688
3.0	**0.776**	**0.596**	**0.691**

Bold represents the maximum value of the column.

4.3. Algorithm Comparison

To demonstrate the effectiveness of our proposed network, we compared it with nine other state-of-the-art (SOTA) methods, including five anchor-based object detection methods (i.e., Faster RCNN, YOLOv5, YOLOv6 [25], YOLOv7, and SSD [26]) and four anchor-free object detection methods (i.e., FCOS [27], YOLOv8, YOLOv9, and DETR [28]). For a fair comparison, all comparison results were generated from the source code provided by the author. All methods were retrained on the same dataset as the approach introduced in this paper, and the original set parameters of the corresponding methods were used.

Table 2 and Figure 6 show the comparison of the four indicators and visualization results of different methods, respectively.

Table 2. Comparison of the MTP-YOLO with other networks. RCNN: region-convolutional neural network. YOLO: You only look once. SSD: Single Shot Multi-Box Detector. FCOS: Fully Convolutional One-Stage object detection. DETR: Detection Transformer. MTP-YOLO: "You only look once"-based Maritime Tiny Person detector.

Methods	Precision	Recall	mAP@0.5	mAP@[0.5,0.9]
Faster RCNN	-	-	0.498	0.211
YOLOv5	0.791	0.583	0.665	0.284
YOLOv6	-	-	0.491	0.228
YOLOv7	0.784	0.628	0.663	0.245
SSD	0.272	0.052	0.057	-
FCOS	-	-	0.581	0.313
YOLOv8	0.758	0.578	0.674	0.315
YOLOv9	0.767	0.597	0.690	0.331
DETR	-	-	0.189	0.050
MTP-YOLO (Ours)	0.776	0.596	**0.691**	**0.331**

Bold represents the maximum value of the column.

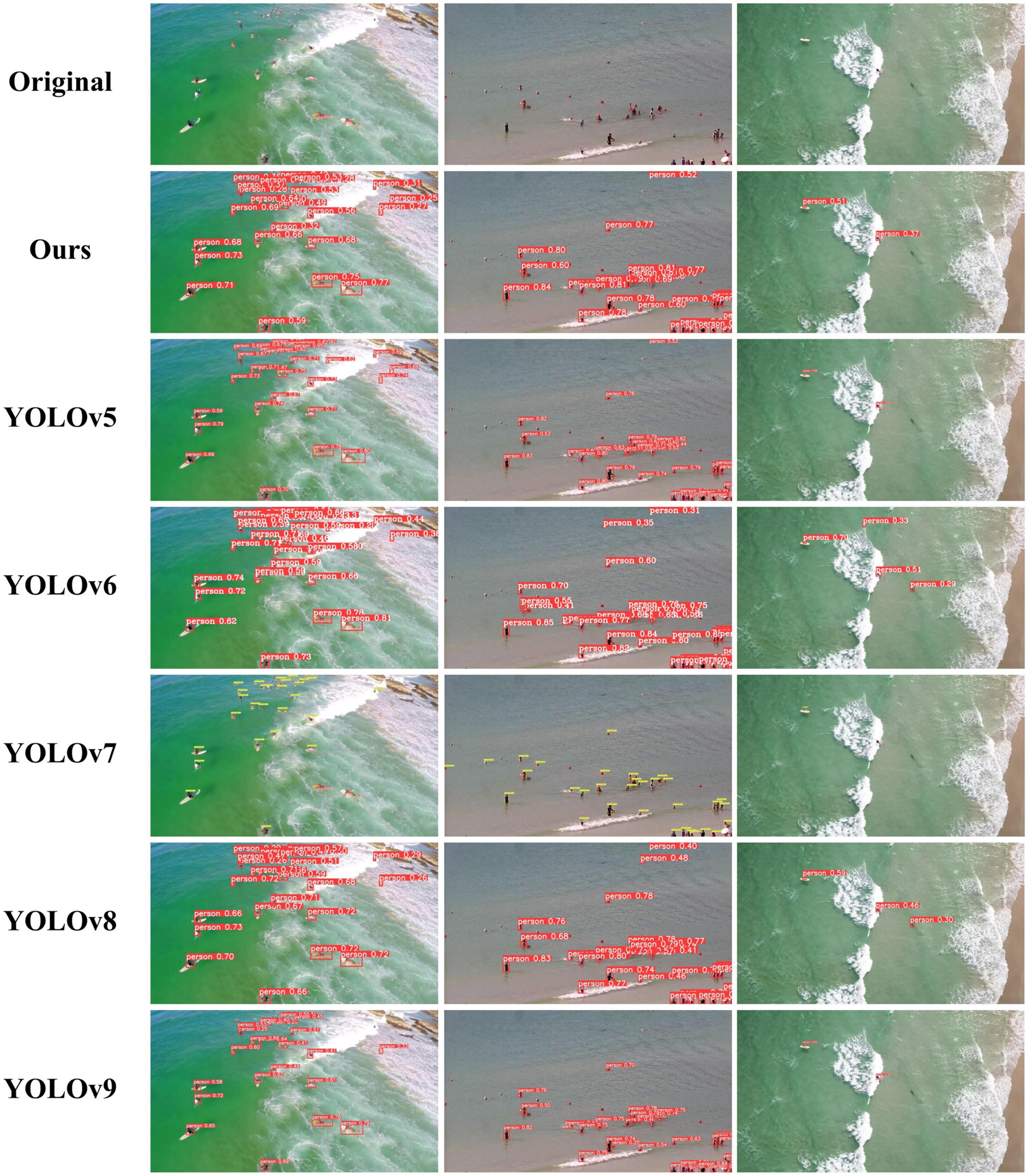


Figure 6. Visual comparison between MTP-YOLO and other networks.

Table 2 shows the evaluation scores of our method and other SOTA methods on precision, recall, and mAP metrics. Among them, precision represents the proportion of true positives among all the positive samples detected, recall represents how many positives in the total sample were predicted correctly, and mAP signifies the mean AP across all classes, where AP denotes the area encompassed by the curve of precision and recall and the coordinate axes. The larger the three indicators, the better. In addition, the parameters of these models are also shown in Table 3. As indicated in Table 2, the proposed MTP-YOLO achieves the highest mAP value, indicating the necessity of a network model specifically designed for detecting tiny persons at sea. For example, compared to the latest YOLOv9 method, our method achieved percentage gains of 0.9% and 0.1% in accuracy and mAP, respectively. In Table 2, the precision and recall of YOLOv7 are better than that of MTP-YOLO as they were obtained at the specified confidence threshold, which indicates that YOLOv7 has a higher precision and recall only at that confidence threshold; however, YOLOv7′s mAP is observed to be lower than that of MTP-YOLO as the mAP is obtained at different confidence thresholds, indicating that MTP-YOLO can better adapt to different confidence thresholds and has high robustness. Furthermore, it should be highlighted that the model size of our method is much smaller than that of the optimal method, YOLOv9. By comparing all indicators, our method achieved a 77.6% accuracy, 56.9% recall, and 69.1% mAP on the TinyPersonv2 dataset using only 48.7% of YOLOv9 parameters. The above results clearly indicate that our model has achieved accuracy comparable to state-of-the-art object detection methods.

Table 3. Comparison of parameters. RCNN: Region Convolutional Neural Network. YOLO: You only look once. SSD: Single Shot Multi-Box Detector. FCOS: Fully Convolutional One-Stage object detection. DETR: Detection Transformer. MTP-YOLO: "You only look once"-based Maritime Tiny Person detector.

Faster RCNN	**YOLOv5**	**YOLOv6**	**YOLOv7**	**SSD**
315.0 M	13.8 M	38.8 M	284.7 M	90.6 M
FCOS	**YOLOv8**	**YOLOv9**	**DETR**	**MTP-YOLO**
244 M	21.5 M	116.7 M	474 M	56.9 M

In Figure 6, we provide a visual (qualitative) comparison to demonstrate the superiority of the proposed MTP-YOLO. The first row comprises the original images, and the second row is our result, followed by YOLOv5, YOLOv6, YOLOv7, YOLOv8, and YOLOv9. Our method can accurately identify the position of tiny persons in the image. Compared with other SOTA methods, our proposed method has a lower missed detection rate and is more suitable for detecting tiny people at sea in emergency rescue missions.

4.4. Ablation Study

To validate the efficiency of our diverse enhancement strategies for the detection of tiny persons, we conducted ablation testing by gradually merging each optimization measure. Table 4 presents the detailed results of these experiments.

Table 4. Results of ablation study. C2fELAN: cross stage partial layer with two convolutions efficient layer aggregation networks. MCE-CBAM: Multi-level Cascaded Enhanced Convolutional Block Attention Module. W-EIoU: Weighted Efficient Intersection over Union.

Methods	Precision	Recall	mAP@0.5
Baseline	0.758	0.578	0.674
Baseline + C2fELAN	0.771	0.597	0.689
Baseline + C2fELAN + MCE-CBAM	0.775	0.596	0.690
Baseline + C2fELAN + MCE-CBAM + W-EIoU	**0.776**	0.596	**0.691**

Bold represents the maximum value of the column.

Analysis of C2fELAN. C2fELAN resulted in an increase in the number of network layers from 225 to 548, and GFLOPS increased from 28.8 G to 78.7 G. However, compared to the baseline, due to the integration of the C2fELAN module in Table 4, the model's precision increased from 0.758 to 0.771, with a percentage gain of 1.30%. The recall rate increased from 0.578 to 0.597, with a percentage gain of 1.90%. In addition, the mAP value of the model has increased from 0.674 to 0.689. The above results indicate the effectiveness of the C2fELAN module.

Analysis of MCE-CBAM. In Table 4, we also demonstrate the effectiveness of the Multi-level Cascaded Enhanced CBAM. In terms of accuracy, this module brought a 0.4% gain to the baseline and contributed 0.1% percentage points to the baseline in terms of mAP. The Multi-level Cascaded Enhanced CBAM enables networks to focus on important regions that are conducive to detecting tiny objects, thereby improving the detection performance of tiny objects.

Analysis of W-EIoU. Table 4 shows that the adoption of the refined Weighted EIoU led to a 0.1% improvement in both the model's accuracy and its mAP score. The experimental results demonstrate that the weighted EIoU loss fully considers the sensitivity of tiny objects to position deviation, and applies larger weights to the position deviation term, making the model more focused on predicting the center point position, thereby improving the recognition ability of tiny objects.

5. Conclusions

This paper proposes an end-to-end object detection network specifically designed for detecting maritime tiny persons, called MTP-YOLO. Benefiting from the proposed C2fELAN feature extraction module, our network can fully capture key features related to tiny objects to accurately locate the objects. We integrated the designed Multi-level Cascaded Enhanced CBAM into our model, improving the capacity of the model to focus on crucial details of tiny objects. In addition, by modifying the bounding box regression loss function to our proposed Weighted EIoU, there is an additional enhancement in the model's capacity to pinpoint the location of tiny objects, leading to a decrease in the rate at which tiny objects go undetected. However, this method is mainly suitable for conditions with good lighting, and its performance may decrease when night approaches. Moving forward, our intention is to obtain images in night or dark scenes through data augmentation or camera shooting, making them suitable for emergency rescue in night scenes, and use model lightweight methods to enhance the detection efficiency of our model, thereby evaluating the practical deployment efficacy of the suggested approach.

Author Contributions: Conceptualization, S.L., Z.Z. and Y.S.; methodology, S.L.; software, S.L.; validation, S.L. and Z.L.; formal analysis, Z.Z. and Y.S.; investigation, S.L. and Z.L.; resources, S.L., Z.Z. and Y.S.; data curation, S.L. and Z.L.; writing—original draft preparation, S.L. and Z.L.; writing—review and editing, Z.Z., Y.S. and X.Z.; visualization, S.L.; supervision, Z.Z., Y.S. and X.Z.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experimental data used for verification in this article can be publicly obtained on the extreme mart, with the identifier: https://www.cvmart.net/dataSets/detail?tabType=1&id=364 (accessed on 26 February 2024).

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

Abbreviations

MTP-YOLO	you only look once based maritime tiny person detector
C2fELAN	cross stage partial layer with two convolutions efficient layer aggregation networks
MCE-CBAM	multi-level cascaded enhanced convolutional block attention module
W-EIoU	weighted efficient intersection over union
GELAN	generalized efficient layer aggregation networks

CBS	conv2d, batch normalization, sigmoid linear unit
SPPF	spatial pyramid pooling fast
DFL	distribution focal loss
BCE	binary cross entropy
RepNC2f	re-parameterization cross stage partial layer with two convolutions without identity connection
RepNBottleneck	re-parameterization bottleneck without identity connection
RepConvN	re-parameterization convolution without identity connection
RepConv	re-parameterization convolution
SiLU	sigmoid linear unit
CAM	channel attention module
SAM	spatial attention module
FC	Fully Connected layer

References

1. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
2. Shehzadi, T.; Hashmi, K.A.; Stricker, D.; Afzal, M.Z. Object Detection with Transformers: A Review. *arXiv* **2023**, arXiv:2306.04670.
3. Chen, G.; Wang, H.; Chen, K.; Li, Z.; Song, Z.; Liu, Y.; Chen, W.; Knoll, A. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 936–953. [CrossRef]
4. Zhou, Z.; Li, Z.; Sun, J.; Xu, L.; Zhou, X. Illumination Adaptive Multi-Scale Water Surface Object Detection with Intrinsic Decomposition Augmentation. *J. Mar. Sci. Eng.* **2023**, *11*, 1485. [CrossRef]
5. Yu, X.; Chen, P.; Wu, D.; Hassan, N.; Li, G.; Yan, J.; Shi, H.; Ye, Q.; Han, Z. Object Localization under Single Coarse Point Supervision. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
6. Zhou, Z.; Hu, X.; Li, Z.; Jing, Z.; Qu, C. A Fusion Algorithm of Object Detection and Tracking for Unmanned Surface Vehicles. *Front. Neurorobot.* **2022**, *16*, 808147. [CrossRef]
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
8. Wang, C.-Y.; Liao, H.Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020.
9. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
10. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017.
12. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO (Version 8.0.0) [Computer Software]. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 26 February 2024).
13. Lim, J.-S.; Astrid, M.; Yoon, H.-J.; Lee, S.-I. Small Object Detection using Context and Attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021.
14. Quan, Y.; Zhang, D.; Zhang, L.; Tang, J. Centralized Feature Pyramid for Object Detection. *IEEE Trans. Image Process.* **2023**, *32*, 4341–4354. [CrossRef]
15. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018.
16. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
17. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA,, 18–23 June 2018.
19. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.

20. 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. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
21. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM International Conference on Multimedia (MM '16), Association for Computing Machinery, New York, NY, USA, 15–19 October 2016.
22. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
23. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
24. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *arXiv* **2021**, arXiv:2101.08158. [CrossRef]
25. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
27. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
28. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872.

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