applied sciences

MDPI

Article

An Object Detection and Localization Method Based on Improved YOLOv5 for the Teleoperated Robot

Zhangyi Chen, Xiaoling Li *, Long Wang, Yueyang Shi, Zhipeng Sun and Wei Sun

School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710000, China
* Correspondence: xjtulxl@mail.xjtu.edu.cn

Abstract: In the traditional teleoperation system, the operator locates the object using the real-time scene information sent back from the robot terminal; however, the localization accuracy is poor and the execution efficiency is low. To address the issues, we propose an object detection and localization method for the teleoperated robot. First, we improved the classic YOLOv5 network model to produce superior object detection performance and named the improved model YOLOv5_Tel. On the basis of the classic YOLOv5 network model, the feature pyramid network was changed to a bidirectional feature pyramid network (BiFPN) network module to achieve the weighted feature fusion mechanism. The coordinate attention (CA) module was added to make the model pay more attention to the features of interest. Furthermore, we pruned the model from the depth and width to make it more lightweight and changed the bounding box regression loss function GIOU to SIOU to speed up model convergence. Then, the YOLOv5_Tel model and ZED2 depth camera were used to achieve object localization based on the binocular stereo vision ranging principle. Finally, we established an object detection platform for the teleoperated robot and created a small dataset to validate the proposed method. The experiment shows that compared with the classic YOLOv5 series network model, the YOLOv5_Tel is higher in accuracy, lighter in weight, and faster in detection speed. The mean average precision (mAP) value of the YOLOv5_Tel increased by 0.8%, 0.9%, and 1.0%, respectively. The model size decreased by 11.1%, 70.0%, and 86.4%, respectively. The inference time decreased by 9.1%, 42.9%, and 58.3%, respectively. The proposed object localization method has a high localization accuracy with an average relative error of only 1.12%.

Keywords: teleoperated robot; object detection; object localization; improved YOLOv5 network; distance estimation

check for updates

Citation: Chen, Z.; Li, X.; Wang, L.; Shi, Y.; Sun, Z.; Sun, W. An Object Detection and Localization Method Based on Improved YOLOv5 for the Teleoperated Robot. *Appl. Sci.* **2022**, *12*, 11441. https://doi.org/10.3390/app122211441

Academic Editor: Franz Wotawa

Received: 26 September 2022
Accepted: 2 November 2022
Published: 11 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The teleoperated robot can help people to complete some complex operation tasks in dangerous or difficult-to-reach environments. By combining human operation experience and judgment ability with robot intelligence, it can overcome the limitations of human physiology and psychology, and environmental geographic space. It greatly expands the human perception ability and operation means. The teleoperated robot is widely applied in demining and detonation, radiation pollution, public health, and other fields [1–6], which can improve work efficiency and prevent operators from being hurt.

In the teleoperation system, the robot is often equipped with a robotic arm to allow the robot to adapt to various complex operations. Given the working space limitation of the robotic arm, precise object localization is required for the teleoperated robot. The object localization accuracy is an important factor that affects the execution efficiency and success rate of teleoperation tasks. In the traditional teleoperation system, during the task execution, the operator mainly relies on the real-time scene information returned from the robot terminal to flexibly adjust the position of the robot, so that the object is in a suitable working space, after which the operator controls the robot to complete the task. Since the object localization largely depends on the operator's experience and on-the-spot judgment, the localization accuracy is low and the execution efficiency is slow.

Computer vision technology can give mechanical systems more powerful and intelligent object recognition capabilities, so it is widely used in robotics, intelligent detection, aerospace, agricultural picking, and other fields [7–12]. At present, the two-stage and one-stage object detection algorithms comprise the majority of the object detection algorithm. Although the latter is greater in terms of detection speed, the former is superior in terms of detection and localization accuracy. The R-CNN [13], SPP-Net [14], Fast R-CNN [15], Faster R-CNN [16], Mask R-CNN [17], and A-Fast-RCNN [18] are the major elements of the two-stage object detection algorithm. The OverFeat [19], YOLO [20], YOLOv2 [21], SSD [22], R-SSD [23], YOLOv3 [24], YOLOv4 [25], and YOLOv5 [26] are the major elements of the one-stage object detection algorithm.

The object detection method is widely applied in robotics because it can rapidly and precisely locate the object within the vision system's field of view, enhancing the system's intelligence and task execution efficiency. In [27], an improved YOLOv3-based litter detection method is proposed for litter-capturing robots, achieving real-time high-precision object detection in dynamic aquatic environments. In [28], an intelligent detection algorithm for seafood is proposed for the underwater robot, which achieves the intelligent detection of underwater objects and guides underwater robots to autonomously grasp seafood. In [29], a high-precision strawberry detection algorithm based on an improved Mask R-CNN is proposed for the fruit picking robot, which solves the problems of poor generality and robustness of traditional object detection algorithms in strawberry detection. However, this research and these applications are mainly oriented to the autonomous grasping robot, and pay more attention to the robustness and accuracy of object detection. In the teleoperated robot, the object localization accuracy and detection speed are important factors that affect task execution efficiency and the operator's interactive experience.

Therefore, we propose an object detection and localization method based on the improved YOLOv5 network model for the teleoperated robot. The following is a summary of our work's main contributions:

(1) Improve the classic YOLOv5 object detection network model; the improved model can effectively increase the accuracy and detection speed of object detection.
(2) Prune the YOLOv5 network model; the resulting lightweight network structure is easier to deploy on embedded devices with insufficient computing power, which is more suitable for the teleoperated robot.
(3) Propose an object localization algorithm based on the improved YOLOv5 network model and the binocular stereo vision ranging principle, which achieves accurate object localization.
(4) Design and establish an object-detection platform for a teleoperated robot, and create a small dataset to validate the proposed method.

The primary organization of this paper is as follows: Section 2 mainly describes an improved YOLOv5 network model and object localization method. Section 3 introduces the experimental design and data acquisition process. Section 4 describes the experimental results and analysis. Section 5 concludes the paper.

2. Method

2.1. Object Detection

2.1.1. YOLOv5 Network Structure

YOLOv5 [26] is a one-stage object detection algorithm with fast detection speed and high detection accuracy. There are four versions of the YOLOv5 algorithm: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, and the network depth and width increase sequentially. The detection accuracy of the four models usually improves sequentially. However, the model size also increases sequentially, making the model difficult to use on embedded devices with insufficient computing power. In addition, as the network parameters increase sequentially, the speed of model training and detection becomes slower, and the possibility of overfitting increases. Therefore, it is very important to balance accuracy, model size, and detection speed for YOLOv5 in practical applications.

The input terminal, the Backbone network, the Neck network, and the head output terminal are the four parts of the classic YOLOv5 network structure. The input terminal contains image preprocessing operations such as scaling the input image to the network input size of 640 × 640 and normalizing it, etc. The Backbone network is a deep convolutional neural network that aggregates and forms image features at different scales. The Focus, Conv, C3, and SPP modules make up the majority of it. The Neck network is a layer that combines and mixes image features, which can further expand the variety of features. The Conv, upsampling, and C3 modules make up the majority of it. The head output terminal generates object bounding boxes, predicts object classes, and predicts results using image features. The classic YOLOv5 network structure is shown in Figure 1.

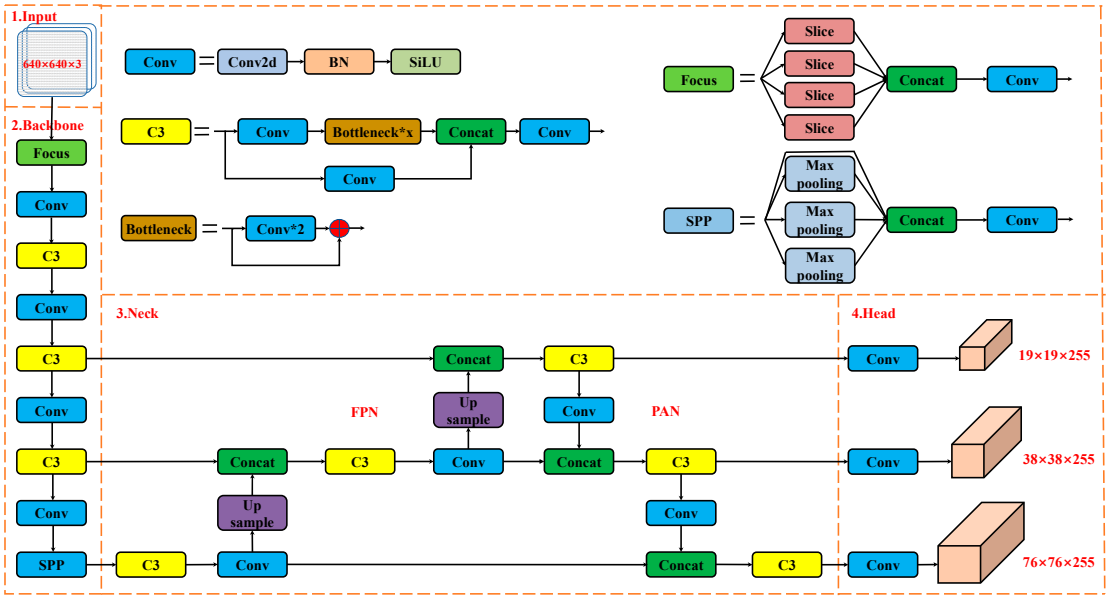

Figure 1. The input terminal, the Backbone network, the Neck network, and the head output terminal are the four parts of the classic YOLOv5 network structure.

2.1.2. Improvement of YOLOv5 Network Structure

Feature Fusion Network Structure Design

In the YOLOv5 model, a feature pyramid module combining feature pyramid network (FPN) [30] and path aggregation network (PAN) [31], is used in the Neck network to achieve deep fusion of feature maps at multiple scales, as shown in Figure 2. Greater location information exists in the bottom-level feature maps, which is helpful for locating objects, whereas stronger semantic data exists in the high-level feature maps, which is helpful for classifying objects. The FPN structure enhances the semantic information of the predicted feature maps, establishing a top-down path, and the low-level features are fused with high-level features through upsampling. The PAN structure improves the location information of the predicted feature maps, which establishes a bottom-up path, and the high-level features are fused with the low-level features through downsampling. The FPN network and the PAN network make the predicted feature maps have both high semantic information and location information, which greatly improves the accuracy of the object detection task.

Although the FPN structure and the PAN structure can fuse multi-scale features, the feature maps are simply superimposed, and all input features are treated equally without distinction when fusing different input feature maps. However, different input features have varying levels of resolution, and they typically make up the output feature inequitably. Tan et al. [32] developed a straightforward and effective bidirectional feature pyramid network (BiFPN) as a solution to this problem, as shown in Figure 2. The BiFPN network repeatedly applies top-down and bottom-up multi-scale feature fusion and learns the importance of different input features using learnable weights to achieve a weighted

feature fusion mechanism. The feature pyramid structure of the Neck network in the classic YOLOv5 model is changed to the BiFPN network structure in our work to improve the feature expression ability of image features and the model's detection accuracy.

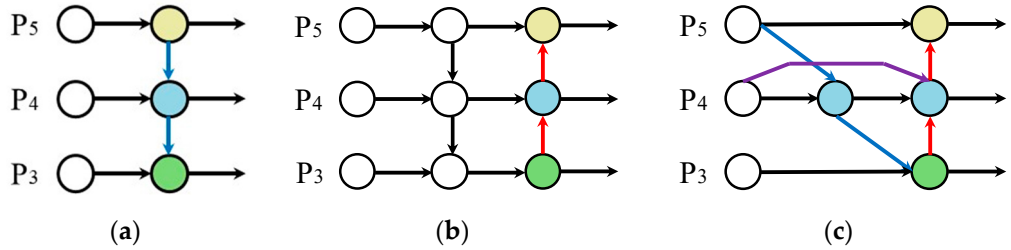

Figure 2. Feature network structure. (**a**) The top-down channel provided by the FPN enables for the fusion of multi-scale features from layers 3 to 5(P3–P5); (**b**) On top of the FPN, the PAN establishes a second bottom-up channel; and (**c**) The BiFPN network with weighted feature fusion and repeated bidirectional cross-scale connections.

Coordinate Attention Module Design

The CA module is a network proposed by Hou et al. [33], as shown in Figure 3. In addition to cross-channel information, it also captures orientation-aware and position-sensitive information, which makes the model accurately locate the exact position of the object of interest and, hence, helps the whole model to recognize better using a small amount of computation.

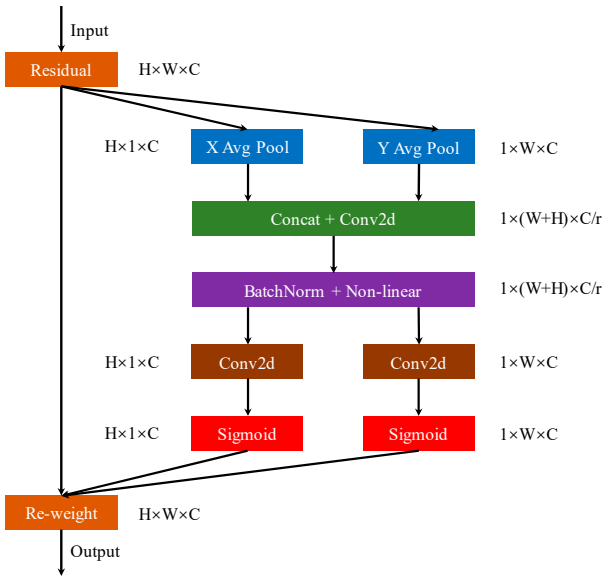

Figure 3. CA module structure.

The CA module divides the multichannel attention into two 1-dimensional feature encoding processes, aggregating features along two different spatial directions. Long-range dependencies in one spatial direction are captured while maintaining exact location information in the other. The generated feature maps are then separately encoded to generate a pair of orientation-aware and position-sensitive feature maps, which can be complementarily applied to the input feature maps $H \times W \times C$, focusing not only on reweighing the importance of different channels, but also on encoding the spatial information.

The specific addition position of the CA module in the YOLOv5_Tel model is shown in Figure 4. With only a small amount of extra computation, the CA module can improve the network model's performance.

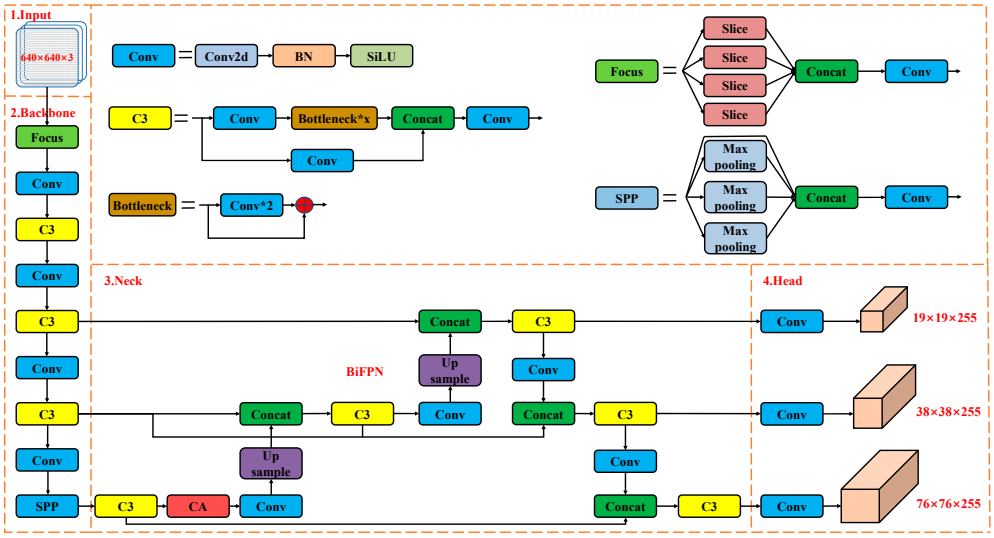

Figure 4. The YOLOv5_Tel model's structure. The feature pyramid module is changed to the BiFPN network and the CA module is added to achieve better detection performance.

Network Pruning

We primarily prune the YOLOv5 network model in terms of network depth and network width to decrease the number of calculations, shorten the network's training and inference time, and make it lightweight so that it may be deployed on embedded devices with insufficient computing power.

We compressed the depth of the network model by controlling the number of Bottleneck residual components in C3 modules by using a hidden layer pruning method. We adjusted the residual components of the eight C3 modules in the Backbone network and the Neck network to 2, 4, 4, 1, 1, 1, 1, 1, respectively. The number of residual components in the C3 modules of different YOLOv5 models is shown in Table 1.

Table 1. Number of residual components in C3 modules in different YOLOv5 models.

Model	Number of Residual Components in C3							
	Backbone			Neck				
	First	Second	Third	First	Second	Third	Fourth	Fifth
YOLOv5s	1	3	3	1	1	1	1	1
YOLOv5m	2	6	6	2	2	2	2	2
YOLOv5l	3	9	9	3	3	3	3	3
YOLOv5_Tel	2	4	4	1	1	1	1	1

We shrink the width of the network model by controlling the number of convolution kernels in Backbone network by using the kernel pruning method. We cut the number of convolution kernels for the Focus module and the four Conv modules to 30, 60, 120, 240, and 480, respectively. The number of convolution kernels in the Focus and Conv modules of different YOLOv5 models is shown in Table 2.

Table 2. Number of convolution kernels in different YOLOv5 models.

Model	Number of Convolution Kernels				
	Focus	First Conv	Second Conv	Third Conv	Fourth Conv
YOLOv5s	32	64	128	256	512
YOLOv5m	48	96	192	384	768
YOLOv5l	64	128	256	512	1024
YOLOv5_Tel	30	60	120	240	480

Loss Function Design

The loss function of the classic YOLOv5 network model is composed of three parts: bounding box regression loss, prediction class loss, and confidence loss. The *GIOU* is used as the bounding box regression loss function, as shown in Equation (1).

$$GIOU = IOU - \frac{|C - (A \cup B)|}{|C|} \tag{1}$$

where $A, B \subseteq S \subseteq R^n$ represent the ground-truth box and predicted box, C represents the minimum circumscribed box of A and B, $C \subseteq S \subseteq R^n$ and $IOU = |A \cap B| / |A \cup B|$.

The *GIOU* [34] function addresses the problem that the *IOU* [35] function is unable to appropriately reflect where the two boxes intersect when the predicted box and the ground truth box do not overlap. The *IOU* function only pays attention to the overlapping area of the ground-truth box and predicted box. The *GIOU* function not only pays attention to the overlapping area, but also pays attention to other non-overlapping areas, which can better reflect the degree of overlap between the ground-truth box and predicted box. However, when the predicted box is completely inside the ground-truth box, the *GIOU* function and the *IOU* function are the same. In this case, the *GIOU* function degenerates into the *IOU* function, which cannot reflect the positional relationship between the ground-truth box and predicted box and causes the bounding box regression slow to converge. Therefore, we changed the loss function of the classic YOLOv5 model from *GIOU* to *SIOU* [36], as shown in Equations (2)–(5).

$$L_{SIOU} = L_{IOU} + L_{ang} + L_{dis} + L_{sha} = 2 - IOU + \frac{\left(1 - e^{-w_w}\right)^{\theta} + \left(1 - e^{-w_h}\right)^{\theta} - e^{-\gamma\rho_x} - e^{-\gamma\rho_y}}{2} \tag{2}$$

$$w_w = \frac{\left|w - w^{gt}\right|}{\max(w, w^{gt})}, w = \frac{\left|h - h^{gt}\right|}{\max(h, h^{gt})} \tag{3}$$

$$\Lambda = 1 - 2\sin^2(\arcsin(\alpha) - \frac{\pi}{4}) \tag{4}$$

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda \tag{5}$$

where (w, h) and (w^{gt}, h^{gt}) represent the width and height of the ground-truth box and predicted box, (b_{c_x}, b_{c_y}) and $(b_{c_x}^{gt}, b_{c_y}^{gt})$ represent the center coordinates of the ground-truth box and predicted box, (c_w, c_h) represents the minimum bounding box width and height of ground-truth box and predicted box, α represents the angle between the horizontal line and the center coordinate line of the ground-truth box and predicted box, and $\theta \in [2, 6]$ is an adjustable parameter to control the degree of attention to the L_{sha}.

The angle loss L_{ang}, the distance loss L_{dis}, the shape loss L_{sha}, and the *IOU* loss L_{IOU} are the four components that make up the *SIOU* function. The *SIOU* function achieves faster convergence in the stage of training and better performance in inference compared to existing methods by introducing directionality in the loss function cost.

The final structure of the YOLOv5_Tel model is shown in Figure 4.

2.2. Object Localization

2.2.1. Binocular Ranging Principle

Accurately estimating the distance between the object and the teleoperated robot is fundamental to object localization. We developed it based on the binocular stereo vision principle. Through parallax, humans are able to estimate an object's distance. Based on this, the binocular stereo vision ranging estimates the distance to an object by the difference between the images captured by the left and right cameras. The geometric model of binocular camera ranging is shown in Figure 5.

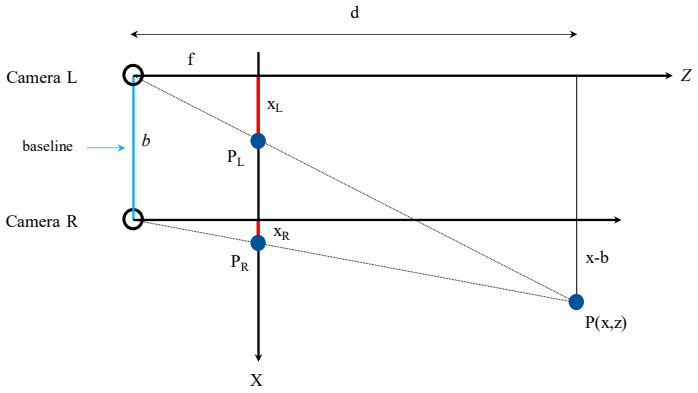

Figure 5. Binocular camera ranging geometric model.

Where *Camera L* and *Camera R* are the optical centers of the left and right cameras, respectively, and the distance between them is called the baseline of the binocular camera, denoted by b. P_L and P_R are the imaging points of the spatial point P on the left and right cameras at the same time, respectively. x_L and x_R are the distances between the imaging points P_L and P_R and the optical axes of the left and right cameras, respectively. f is the focal length. d is the distance between the spatial point p and the baseline of the binocular camera.

The intrinsic and extrinsic parameters of the left and right cameras are the same, in theory. The imaging points P_L and P_R also differ only on the x-axis because the optical centers of the left and right cameras only have a positional deviation at the x-axis. According to the law of trigonometric similarity, the distance d can be solved through geometric relations, as shown in Equation (6).

$$d = \frac{f * b}{x_L - x_R} = \frac{f * b}{u} \tag{6}$$

where u is the difference between the x-axis coordinates of the imaging points P_L and P_R, which is called parallax.

The focal length f and the baseline distance b of the binocular camera can be obtained through camera calibration. Therefore, after calculating the parallax u of the target point in the left and right cameras, the distance between the target point and the camera can be calculated through Equation (6).

2.2.2. Distance Estimation

In our work, we use the ZED2 depth camera and the YOLOv5_Tel model to achieve real-time distance estimation between the object and the camera left eye. The ZED2 depth camera can obtain the 3D point cloud coordinates (x, y, z) of any pixel in the image based on binocular stereo vision. The object class, the confidence score, and the predicted box coordinate (x_{min}, y_{min}, x_{max}, y_{max}) are returned when the object is detected by the YOLOv5 network model, where (x_{min}, y_{min}) is the coordinate of the top left-most of the predicted box and (x_{max}, y_{max}) is the coordinate of the down right-most of the predicted box. It is possible to determine an object's pixel point coordinates within an image as well as the matching three-dimensional point cloud coordinates when an object is detected. The distance between the object and the left eye of the camera can be calculated through the Euclidean distance formula, as shown in Equation (7).

$$distance = \sqrt{x^2 + y^2 + z^2} \tag{7}$$

Researchers usually only solve the distance between the center point of the object and the left eye of the camera as the distance estimation result [37]. However, the depth information of some pixels cannot be obtained due to occlusion or anomaly in the standard

depth sensing mode of the ZED2. If the depth information of the center point cannot be exactly obtained, it will cause the abnormal distance estimation of the object. In our work, we used the center point of the prediction box and its surrounding 8 points to solve the distance estimation to improve the accuracy and the robustness to lighting conditions and occlusions, and their position distribution is shown in Figure 6.

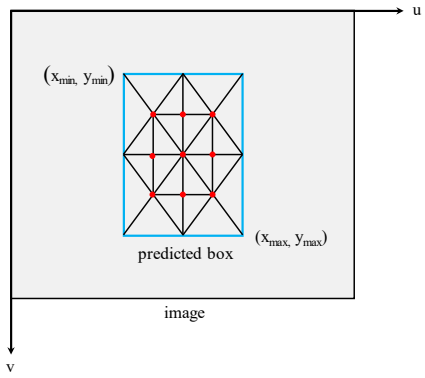

Figure 6. The positional relationship of the 9 points used for distance estimation as indicated by the red dots.

In the teleoperation system, when the distance is too close or too far between the robot and the object, the operation task cannot be done so that there is no need for distance estimation to the object. In addition, the predicted box and label information of the object are prone to occlusion in the image, which is not conducive to the feedback of environmental information. Therefore, the object is detected and estimated distance in the range of 0.3–3 m between the object and the left eye of the camera in our work. The algorithm of the distance estimation method is shown in Algorithm 1.

Algorithm 1 Object Detection and Distance Estimation

Require: Self-trained YOLOv5 model and ZED2 depth camera
Ensure: Objects class and their distance from the camera's left eye
Initialize the depth camera(HD720p, 60FPS, depth minimum distance = 0.3 m)
Load the self-trained YOLOv5 model
While(**true**)
    Stereo image is captured for each frame (3D image)
    The image is pre-processed and resized to (640 × 640)
    Input the image into the self-trained YOLOv5 model for object detection and return (x_{min}, y_{min}, x_{max}, y_{max})
    for each *object* **do**
        if (*object* == 1) **then**
            Calculate the coordinates of 9 points
            distance = 0, *n* = 0
            for each $point_i$ **do**
                if ($point_i$.*isValidMeasure*()) **then**
                    $dis_i = \sqrt{x_i^2+y_i^2+z_i^2}$
                    $distance = (distance * n + dis_i)/(n + 1)$
                    $n = n + 1$
                end if
            end for
            if ($0.3 \leq distance$ and $distance \leq 3$) **then**
                Plot predicted box, class, confidence, distance labels
            end if
        end if
    end for

3. Experimental Design

3.1. Dataset

A small dataset was created for the teleoperated robot grasping task to simulate the performance of the different YOLOv5 network models. The dataset contained 14 objects, as shown in Figure 7. A total of 800 images were collected with the ZED2 camera at different angles and position distributions under natural lighting conditions. During the image collection process, we did not deliberately change the lighting conditions, nor did we deliberately place the objects. The objects may have occluded each other, which is more in line with the real working environment of the teleoperate robot. By flipping, scaling, rotating, and cropping the images, we expanded the dataset's size [38]. After that, the dataset contained a total of 4000 images, of which 800 were captured using the ZED2 camera and 3200 were obtained through data augmentation. Finally, the images were labeled by labelImg and inputted into the different models to simulate.

Figure 7. The introduction to the dataset. (**a**) 14 objects in the dataset. The ground truth labels from 1 to 14 are chewing_gum, scissors, snickers, screwdriver, tape, wooden_ball, beer, cylinder, biscuit, grenade, flashbang, milk, landmine, and smoke_bomb; (**b**) The collection environment of the dataset.

3.2. Simulation

Different YOLOv5 models were simulated using Pytorch under the Ubuntu16.04 operating system based on NVIDIA Jeston AGX Xavier, 512-core Volta GPU with Tensor Core, and 8-core ARM64 CPU.

The hyperparameters batch_size and learning rate are crucial to the model's performance. We therefore optimized some hyperparameters based on experiments to further improve the model performance. In the simulation, the batch_size was set to 64, the learning rate was set to 0.001, the number of iteration epoch was set to 1500, the input image size was set to 640 × 640, and the stochastic gradient descent strategy (SGD) was used for parameter update to speed up the convergence. In addition, other parameters used default values. A total of 4000 images made up the dataset, of which 3200 were used for training and 800 for testing.

3.3. Distance Estimation

We designed and established an object detection platform for a teleoperated robot to test the accuracy of the distance estimation method. The teleoperated robot used for the experiment is shown in Figure 8.

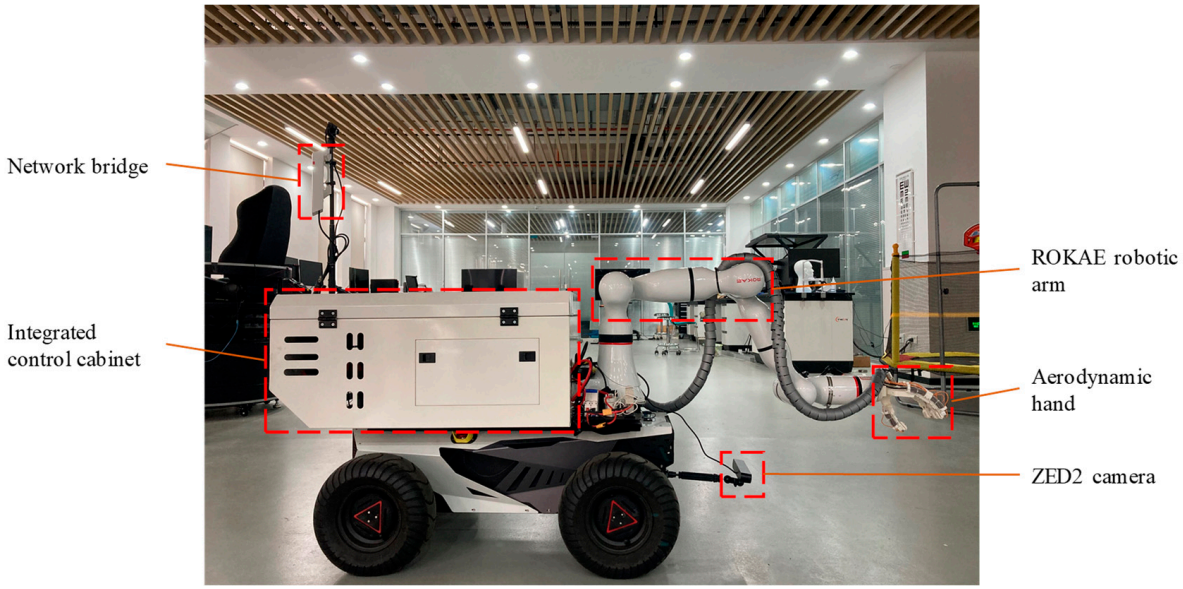

Figure 8. The teleoperated robot experiment platform for distance estimation.

In the experiment, the depth mode of the ZED2 camera was set to ULTRA, the minimum depth distance was set to 0.3 m, the resolution was set to HD720P, and the frame rate was set to 60. In addition, other parameters used default values. In the range of 0.3–3 m, the interval was divided into 9 sub-intervals with a step size of 0.3 m, and the distance estimation accuracy in each sub-interval was tested. The object was randomly placed on the ground in the experiment. The actual distance between the object and the left eye of the ZED2 camera was obtained by manual measurement, and the average of three measurements was taken as the final result. We conducted 3 rounds of experiments on the 14 objects in the dataset in the 9 sub-intervals, and a total of $14 \times 9 \times 3 = 378$ valid data was collected.

4. Result and Analysis

4.1. Simulation Results and Analysis

In the simulation, we trained 1500 epochs on the classic YOLOv5s, YOLOv5m, YOLOv5l, and the YOLOv5_Tel. The loss function convergence curves of the four network models on the training set are shown in Figure 9.

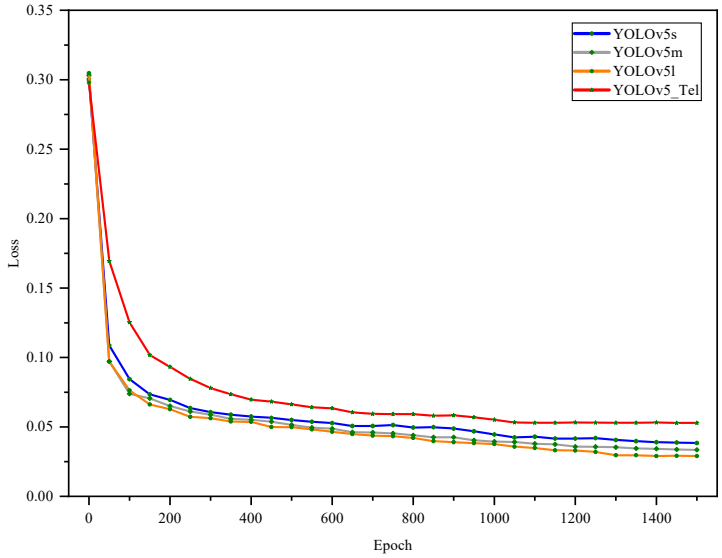

Figure 9. Convergence of loss functions for different models on the training set.

To assess the effectiveness of the YOLOv5_Tel model, the simulation results of the four network models were compared and analyzed. The evaluation metrics for model detection accuracy are precision, recall, mAP, and F1-score [12]. The model's degree of lightweight and detection speed was characterized using the model size and inference time. The simulation results of the four different models are shown in Table 3.

Table 3. The simulation results of different YOLOv5 models.

Models	Precision (%)	Recall (%)	mAP (%)	F1 (%)	Model Size (MB)	Inference Time (ms)
YOLOv5s	91.7	98.6	98.6	95.0	14.4	22
YOLOv5m	91.5	98.5	98.5	94.9	42.6	35
YOLOv5l	91.4	98.4	98.4	94.8	93.8	48
YOLOv5_Tel	92.6	99.4	99.4	95.9	12.8	20

The comparison and analysis of the YOLOv5_Tel model and the classic model was done in light of the results of the simulation.

(1) In the training stage, the loss function value of the YOLOv5_Tel model tends to stabilize after 1100 iterations, whereas the classic YOLOv5s, YOLOv5m, and YOLOv5l models tend to stabilize after 1400 iterations, indicating that the YOLOv5_Tel model converges faster, which can reduce the training cost.

(2) The YOLOv5_Tel model offers better object detection accuracy, as demonstrated by higher precision, recall, mAP value, and F1-score of 92.6%, 99.4%, 99.4%, and 95.9%, respectively.

(3) The YOLOv5_Tel model is 12.8MB in model size, which is smaller than the classic YOLOv5s, YOLOv5m, and YOLOv5l network models by 11.1%, 70.0%, and 86.4%, respectively, indicating that the YOLOv5_Tel is lighter and better suited for embedded devices with insufficient computing power.

(4) The YOLOv5_Tel model outperforms the classic YOLOv5s, YOLOv5m, and YOLOv5l network models in terms of inference time by 9.1%, 42.9%, and 58.3%, respectively, indicating that the YOLOv5_Tel has a faster detection speed and is more suitable for the teleoperated robot.

4.2. Distance Estimation Results and Analysis

In the actual scene, objects were haphazardly placed on the ground between 0.3 and 3 m from the camera. The effect of the object detection and localization method proposed in this paper is shown in Figure 10.

The proposed object detection method can accurately and robustly detect the objects. The proposed object localization method can estimate the distance between the object and the camera in real time. It makes it easy for even a novice operator to control the distance between the teleoperated robot and the object, so that the object is in the working space of the robotic arm to achieve fast and accurate localization, which can improve the execution efficiency of the teleoperated robot.

To further evaluate the accuracy of distance estimation, we calculated the relative error between the estimated distance and the true distance for 378 valid distance estimation data. The relative distance error of all objects was averaged to characterize the accuracy of the distance estimation method in different sub-intervals. The distance estimation accuracy results are shown in Figure 11.

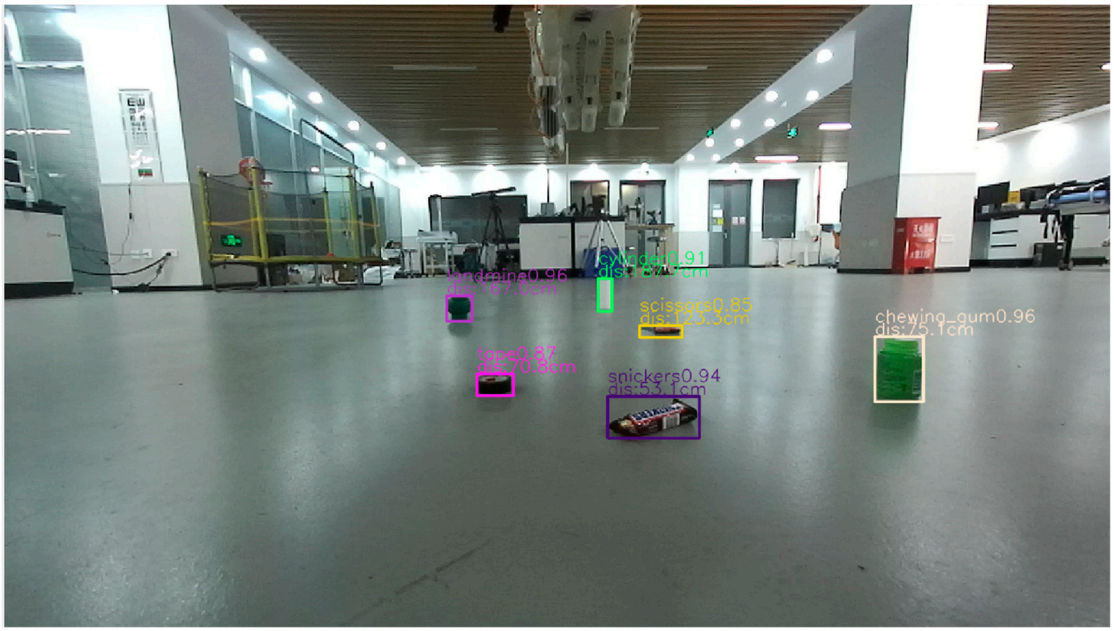

Figure 10. Actual effect of object detection and localization.

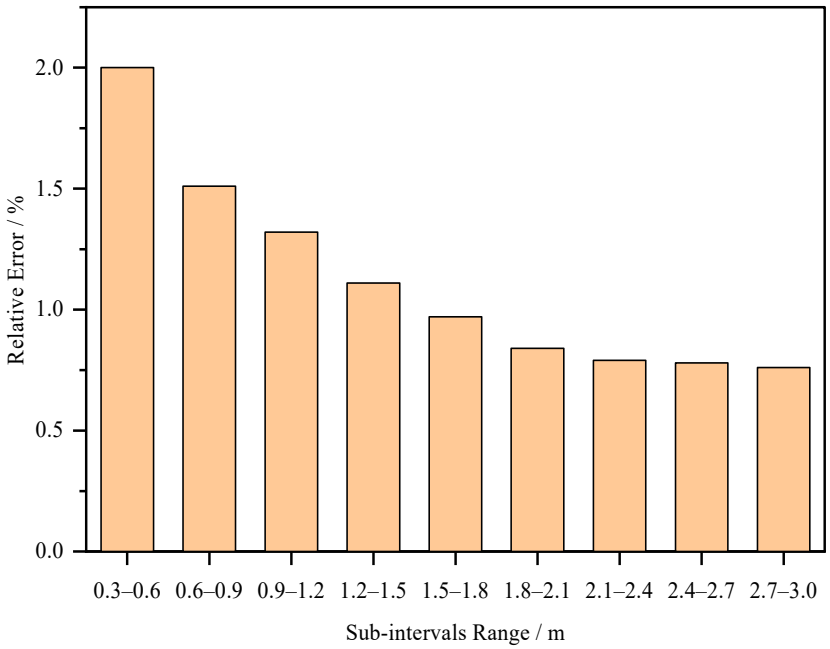

Figure 11. Relative error of distance estimation in different sub-intervals.

The proposed distance estimation method can achieve good localization accuracy, with a minimum relative error of 0.76%, a maximum relative error of 2.0%, and an average relative error is 1.12%. The proposed distance estimation method is robust within the experimental range, according to the root mean square error (RMSE) of the relative error, which is 0.4%.

5. Conclusions

We propose an object detection and localization method based on improved YOLOv5 for the teleoperated robot that can offer accurate and fast object detection and localization. Based on the outcomes of the experiment, the following conclusions can be drawn:

(1) The model simulation experiment shows that the YOLOv5_Tel model is more accurate, lighter, and faster in object detection. The YOLOv5_Tel model's precision, recall, mAP value, and F1 score are 92.6%, 99.4%, 99.4%, and 95.9%, respectively. The YOLOv5_Tel model's mAP value increased in comparison to the classic YOLOv5s, YOLOv5m, and YOLOv5l models by 0.8%, 0.9%, and 1.0%, respectively. The model size decreased by 11.1%, 70.0%, and 86.4%, respectively, whereas the inference time decreased by 9.1%, 42.9%, and 58.3%.

(2) The distance estimation experiment shows that the object localization method has good localization accuracy and robustness, with an average relative error of distance estimation of 1.12% and a RMSE of relative error of 0.4%.

Although the improved YOLOv5 model performs well in object detection and localization for the teleoperated robot, the model's performance in actual applications is still constrained by the complexity of the working environment. Future work will improve the model's detection robustness and adapt it to more challenging environments.

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

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Adamides, G.; Katsanos, C.; Parmet, Y.; Christou, G.; Xenos, M.; Hadzilacos, T.; Edan, Y. HRI usability evaluation of interaction modes for a teleoperated agricultural robotic sprayer. *Appl. Ergon.* **2017**, *62*, 237–246. [CrossRef] [PubMed]
2. Qian, K.; Song, A.; Bao, J.; Zhang, H. Small Teleoperated Robot for Nuclear Radiation and Chemical Leak Detection. *Int. J. Adv. Robot. Syst.* **2012**, *9*, 70. [CrossRef]
3. Rahman, M.M.; Balakuntala, M.V.; Gonzalez, G.; Agarwal, M.; Kaur, U.; Venkatesh, V.L.N.; Sanchez-Tamayo, N.; Xue, Y.; Voyles, R.M.; Aggarwal, V.; et al. SARTRES: A semi-autonomous robot teleoperation environment for surgery. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2020**, *9*, 376–383. [CrossRef]
4. Novák, P.; Kot, T.; Babjak, J.; Konečný, Z.; Moczulski, W.; Rodriguez López, Á. Implementation of Explosion Safety Regulations in Design of a Mobile Robot for Coal Mines. *Appl. Sci.* **2018**, *8*, 2300. [CrossRef]
5. Koh, K.H.; Farhan, M.; Yeung, K.P.C.; Tang, D.C.H.; Lau, M.P.Y.; Cheung, P.K.; Lai, K.W.C. Teleoperated service robotic system for on-site surface rust removal and protection of high-rise exterior gas pipes. *Autom. Constr.* **2021**, *125*, 103609. [CrossRef]
6. Lin, Z.; Gao, A.; Ai, X.; Gao, H.; Fu, Y.; Chen, W.; Yang, G.-Z. ARei: Augmented-Reality-Assisted Touchless Teleoperated Robot for Endoluminal Intervention. *IEEE/ASME Trans. Mechatron.* **2021**, *27*, 1–11. [CrossRef]
7. Liu, M.; Wang, X.; Zhou, A.; Fu, X.; Ma, Y.; Piao, C. UAV-YOLO: Small Object Detection on Unmanned Aerial Vehicle Perspective. *Sensors* **2020**, *20*, 2238. [CrossRef]
8. Zhang, H.; Li, D.; Ji, Y.; Zhou, H.; Wu, W.; Liu, K. Toward New Retail: A Benchmark Dataset for Smart Unmanned Vending Machines. *IEEE Trans. Ind. Inform.* **2020**, *16*, 7722–7731. [CrossRef]
9. Xue, J.; Zheng, Y.; Dong-Ye, C.; Wang, P.; Yasir, M. Improved YOLOv5 network method for remote sensing image-based ground objects recognition. *Soft Comput.* **2022**, *26*, 10879–10889. [CrossRef]
10. Wang, J.; Gao, Z.; Zhang, Y.; Zhou, J.; Wu, J.; Li, P. Real-Time Detection and Location of Potted Flowers Based on a ZED Camera and a YOLO V4-Tiny Deep Learning Algorithm. *Horticulturae* **2021**, *8*, 21. [CrossRef]
11. Lin, S.Y.; Li, H.Y. Integrated Circuit Board Object Detection and Image Augmentation Fusion Model Based on YOLO. *Front. Neurorobot.* **2021**, *15*, 762702. [CrossRef] [PubMed]
12. Chen, Z.; Wu, R.; Lin, Y.; Li, C.; Chen, S.; Yuan, Z.; Chen, S.; Zou, X. Plant Disease Recognition Model Based on Improved YOLOv5. *Agronomy* **2022**, *12*, 365. [CrossRef]
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]

15. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Ren, S.Q.; He, K.M.; 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] [PubMed]
17. He, K.M.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef] [PubMed]
18. Wang, X.L.; Shrivastava, A.; Gupta, A. A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3039–3048.
19. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 779–788.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
22. 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; pp. 21–37.
23. Jeong, J.; Park, H.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. *arXiv* **2017**, arXiv:1705.09587.
24. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
26. Jocher, G.; Stoken, A.; Borovec, J.; NanoCode012; Chaurasia, A.; Liu, C.; Xie, T.; Abhiram, V.; Laughing; Tkianai; et al. Ultralytics/yolov5: v5.0-YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. *Zenodo* **2021**, *2021*, 4679653. [CrossRef]
27. Li, X.; Tian, M.; Kong, S.; Wu, L.; Yu, J. A modified YOLOv3 detection method for vision-based water surface garbage capture robot. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420932715. [CrossRef]
28. Xu, F.; Dong, P.; Wang, H.; Fu, X. Intelligent detection and autonomous capture system of seafood based on underwater robot. *J. Beijing Univ. Aeronaut. Astronaut.* **2019**, *45*, 2393–2402. [CrossRef]
29. Yu, Y.; Zhang, K.; Yang, L.; Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
30. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
32. Mingxing, T.; Ruoming, P.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
33. Qibin, H.; Daquan, Z.; Jiashi, F. 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; pp. 13708–13717. [CrossRef]
34. Rezatofighi, H.; Tsoi, N.; JunYoung, G.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666. [CrossRef]
35. 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, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
36. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
37. Manjari, K.; Verma, M.; Singal, G.; Kumar, N. QAOVDetect: A Novel Syllogistic Model with Quantized and Anchor Optimized Approach to Assist Visually Impaired for Animal Detection using 3D Vision. *Cogn. Comput.* **2022**, *14*, 1269–1286. [CrossRef]
38. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]