World Electric Vehicle Journal

MDPI

Article

Research on YOLOv5 Vehicle Detection and Positioning System Based on Binocular Vision

Yixiao Zhang [1], Yuanming Gong [1,*] and Xiaolong Chen [2]

[1] School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; m310121460@sues.edu.cn
[2] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; cxllxj@sues.edu.cn
* Correspondence: 06090010@sues.edu.cn

Abstract: Vehicle detection and location is one of the key sensing tasks of automatic driving systems. Traditional detection methods are easily affected by illumination, occlusion and scale changes in complex scenes, which limits the accuracy and robustness of detection. In order to solve these problems, this paper proposes a vehicle detection and location method for YOLOv5(You Only Look Once version 5) based on binocular vision. Binocular vision uses two cameras to obtain images from different angles at the same time. By calculating the difference between the two images, more accurate depth information can be obtained. The YOLOv5 algorithm is improved by adding the CBAM attention mechanism and replacing the loss function to improve target detection. Combining these two techniques can achieve accurate detection and localization of vehicles in 3D space. The method utilizes the depth information of binocular images and the improved YOLOv5 target detection algorithm to achieve accurate detection and localization of vehicles in front. Experimental results show that the method has high accuracy and robustness for vehicle detection and localization tasks.

Keywords: binocular vision; YOLOv5 algorithm; stereo matching; vehicle detection; ranging; positioning

Citation: Zhang, Y.; Gong, Y.; Chen, X. Research on YOLOv5 Vehicle Detection and Positioning System Based on Binocular Vision. *World Electr. Veh. J.* **2024**, *15*, 62. https://doi.org/10.3390/wevj15020062

Academic Editor: Joeri Van Mierlo

Received: 21 December 2023
Revised: 1 February 2024
Accepted: 8 February 2024
Published: 11 February 2024

1. Introduction

The rapid development of artificial intelligence has broadened the development space for intelligent transportation systems and autonomous driving technology [1]. In these applications, accurate vehicle detection and location is a crucial task, which is of great significance for realizing safe driving, improving traffic flow efficiency and improving the driving experience [2]. In recent years, target detection algorithms based on deep learning, such as R-CNN, YOLO and SSD [3,4], have made a breakthrough in the field of computer vision, with the advantages of high efficiency and accuracy. However, monocular cameras have a limited viewing angle, which limits the accuracy of vehicle detection and positioning [5]. To overcome this problem, the use of binocular vision systems can provide more image information and spatial depth information, which can better solve the problem of the accurate localization and detection of vehicles. Binocular vision systems simulate human binocular observation by calculating the pixel displacement (i.e., the parallax) between the two cameras [6,7]. The depth of an object in 3D space can be estimated, which provides a more accurate basis for vehicle detection and localization [8].

With the continuous development and advancement of deep learning technology in recent years, its application in detection and localization tasks has become more and more widespread. Zhang et al. [9] proposed a binocular vision detection and localization method for ships based on the YOLOv3 algorithm, which uses the K-mean clustering method to generate aiming point frames to improve the performance of ship detection and combines with the binocular ranging algorithm to achieve the localization of the ship. Chen et al. [10] proposed an improved algorithm based on YOLOv5-OBB, which chemicalizes

the backbone module, introduces the GELU activation function and adds the CA attention mechanism to enhance the feature representation capability. The improved algorithm increases the accuracy of vehicle position and tilt angle detection to reduce the computation of the model and increase the speed of vehicle position detection. Niu et al. [11] made innovative modifications based on the YOLOv7 framework: the SE attention mechanism was added to the backbone module, the related module was replaced, the feature extraction of the model was enhanced, the model achieved better results and the detection capability was improved. Shao et al. [12] proposed an improved YOLOv5s vehicle recognition and detection algorithm using the ELU activation function instead of the original activation function. The attention mechanism module is added to the backbone network of the YOLOv5s algorithm to improve the feature extraction of small and medium-sized targets. The CIOU loss function replaces the original regression function of the YOLOv5s, which enhances the convergence rate of the loss function and the measurement accuracy. Hu et al. [13] proposed a real-time detection algorithm based on the improved YOLOv5. A new gradient path is set on the spatial pyramid pooling to strengthen the feature extraction capability; learnable adaptive weights for deep and shallow convolutional features are added to the neck feature fusion to better fuse deep semantic and shallow detail features and improve the detection accuracy of small targets. Yao et al. [14] proposed a color-order offset compensation model, optimized the backbone parameters of the YOLOv5 algorithm, fused the attention mechanism and improved the loss function. The improved algorithm increased the detection rate of the target. Zhang et al. [15] improved the YOLOv5 algorithm by adding a lightweight convolutional attention module to the model and introducing a spatial pyramid pooling structure, which improved the detection accuracy of the algorithm. They also used the spatial geometric relationship of the pixels in their images to achieve spatial localization of aircraft hatches.

In the automatic driving environment perception task, it is especially important to realize accurate vehicle detection and spatial location determination. This paper proposes a vehicle detection and localization system based on binocular vision and deep learning. By improving the YOLOv5 algorithm and adding the CBAM attention mechanism, the system enhances its ability to pay attention to vehicle features. Additionally, the original CIOU loss function is replaced with the more accurate EIOU metric, which improves the degree of match between the prediction frame and the real frame, enhancing the target detection ability. The improved YOLOv5 algorithm is combined with the depth information acquired by binocular vision, enabling accurate vehicle detection and localization. Firstly, vehicle images are obtained using the binocular camera, and the 3D spatial information of the vehicle is obtained through a depth estimation algorithm. Then, based on the improved YOLOv5 target detection algorithm, the vehicle is detected in the image information to achieve fast and accurate target recognition. Finally, accurate vehicle localization is achieved by combining the binocular depth information and the improved YOLOv5 detection results.

This paper explores the fusion of binocular vision and improved YOLOv5 target detection, deploying the application on an embedded device capable of simultaneously realizing vehicle detection and 3D spatial position calculation. The method is able to achieve high-precision vehicle detection and localization in road environments with high computational efficiency, through which we can overcome the limitations of traditional methods, improve the accuracy and real-time performance of detection and localization and provide both more reliable technical support for application in the field of intelligent transportation and an effective solution for the application of intelligent transportation systems and automatic driving technology.

2. YOLOv5 Target Detection Algorithm

The YOLO series of target detection algorithms is a series of single-stage deep learning target detectors. Its basic principle is to decompose the image into a grid system, and each grid is responsible for detecting the targets in the area [16]. YOLOv5 has high detection speed and accuracy. The algorithm adopts a convolutional neural network structure

similar to a CNN, but by optimizing the network through techniques such as pruning and quantization, it achieves faster detection speed and lower calculation costs [17].

2.1. YOLOv5 Algorithm Principle

YOLOv5 has improved the model architecture and training strategy and has faster and more accurate detection performance. The input image is divided into a set of non-overlapping grids. Each grid is predicted to obtain candidate bounding boxes [16]. These bounding boxes are represented by central coordinates (width and height), and the probability of the object category to which each bounding box belongs is also estimated. The NMS algorithm is used to screen candidate boxes with more overlap, and the bounding box with the highest probability is selected [18,19]. Finally, the filtered bounding box is output, and the object category and position information are given. The basic YOLOv5 structure is shown in Figure 1.

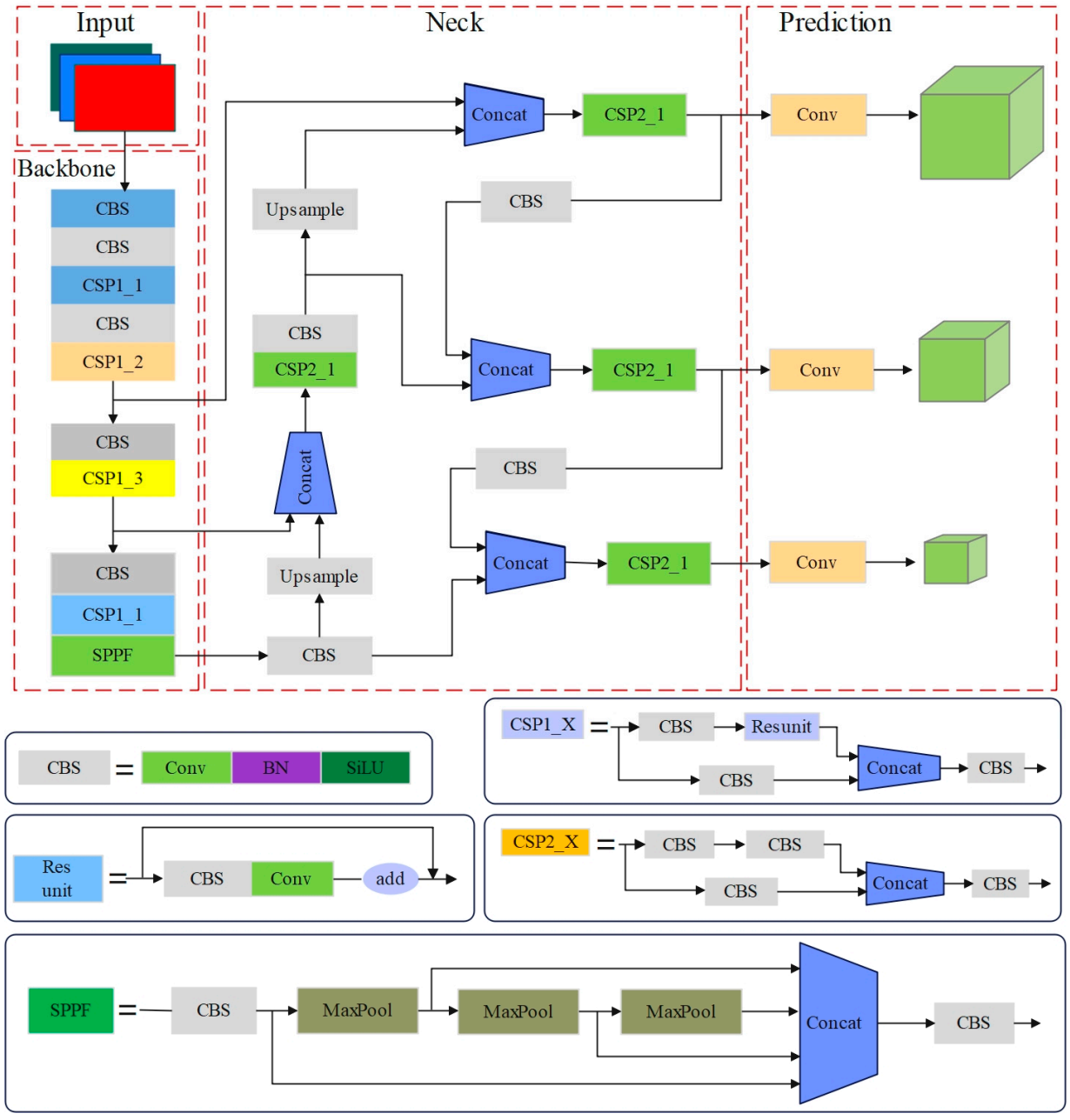

Figure 1. YOLOv5 network structure diagram.

The YOLOv5 model consists of four main parts: the input end, the backbone network, the neck network and the detection head [20].

The input terminal is mainly responsible for pre-processing input pictures so that the network can handle them better. The processed images are sent to the backbone network for feature extraction. The backbone network consists of several key modules, including the Focus module, the convolution module, the CSP module and the spatial pyramid pooling fusion (SPPF) module. The function of these modules is to extract the features of the input images. Firstly, the high-resolution feature map is specially downsampled by the Focus operation, and several low-resolution feature maps are obtained. Then, the processed image is processed through the CBS layer. Finally, through the spatial pyramid pooling layer, including convolution, maximum pooling layer and mosaic fusion, the features are further extracted. The neck network is used to integrate the features output by the backbone network and provide them to the detection head. Its main function is to mix and combine the extracted features and transmit them to the prediction layer, thus shortening the information transmission path, maintaining high semantic information and enhancing the ability of network feature fusion.

2.2. Improved YOLOv5 Algorithm

2.2.1. Adding Attention Mechanisms

In view of the interference of background information with the detection target, this section adds a lightweight convolution attention module (CBAM) and attention mechanism to the detection head structure of YOLOv5, which can strengthen attention to the target during detection, reduce the interference influence of complex environments, improve the target detection effect under background interference and improve detection accuracy. The specific network structure is shown in Figure 2.

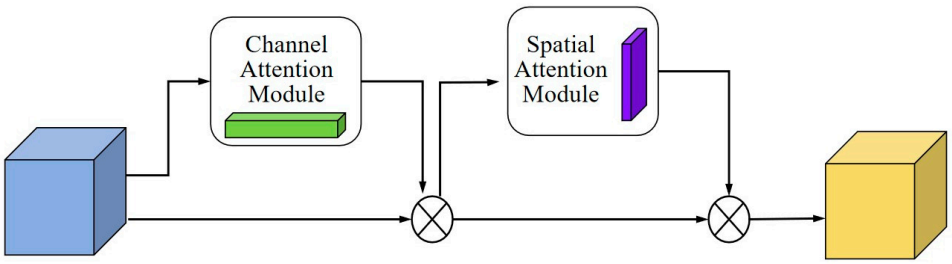

Figure 2. CBAM module structure.

The CBAM attention mechanism can model both channel attention and spatial attention at the same time, so that the model can adaptively select important characteristic channels and spatial positions [21]. The basic principle of CBAM's attention structure is the following: as can be clearly seen from the above figure, CBAM consists of two independent sub-modules, namely, the channel attention module (CAM) and the spatial attention module (SAM), which respectively fuse the attention features in the channel and the spatial dimensions.

As shown in Figure 3, the channel attention module is used to weight different channels of the input feature map to highlight important channel information. It calculates the global average and global maximum of each channel and generates a channel attention vector by using the fully connected layer and activation function. Then, the vector is multiplied by the input feature map using a multiplication operation, thus highlighting the important channel features.

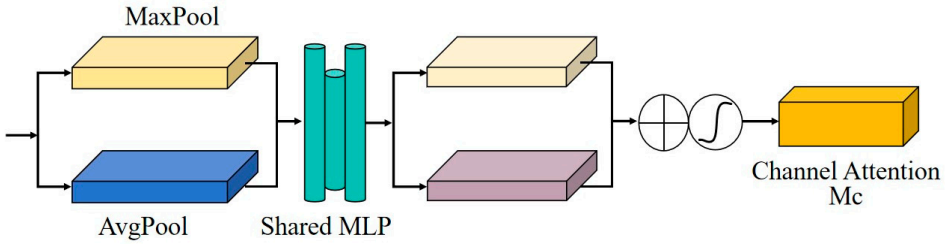

Figure 3. Channel attention module.

As shown in Figure 4, the spatial attention module is used to weight each spatial position to highlight important spatial information. It obtains two feature maps by convolution operation on the input feature map in the spatial dimension, one for calculating the average value of the position and the other for calculating the maximum value of the position. These two feature maps are cascaded and processed by a full connection layer and activation function to generate a spatial attention map. Then, the spatial attention tries to multiply the input feature map using a multiplication operation, thus highlighting the important spatial features.

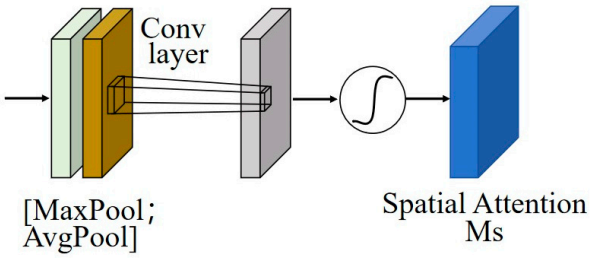

Figure 4. Spatial attention module.

In the process of using the improved YOLOv5 algorithm to detect the target vehicle, the introduction of the CBAM attention mechanism can enhance the ability to pay attention to the vehicle area, help the model to pay more attention to the vehicle area and improve the performance of vehicle detection.

2.2.2. Loss Function Optimization

The loss function evaluates the match between the model predictions and the actual data in YOLOv5 and consists of three components: border regression, confidence and categorical probability. They calculate the deviation of the predicted box from the actual box, the difference between the predicted confidence and the actual confidence and the gap between the predicted category probability and the actual category probability, respectively. Choosing an appropriate loss function can speed up model convergence and improve accuracy and performance. The YOLOv5 model uses the CIOU loss function by default to calculate the localization loss [22]. This function not only considers the overlap area between borders and the distance from the center point but also takes the aspect ratio into account when performing the bounding box regression calculation, which is shown in Equation (1).

$$L_{CIOU} = 1 - IOU + \frac{\rho^2(a, b^{gt})}{c^2} + av \tag{1}$$

$$IOU = \frac{A \cap B}{A \cup B} \tag{2}$$

$$a = \frac{v}{(1 - IOU) + v} \tag{3}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{4}$$

where IOU denotes the ratio of intersection and concatenation between the predicted and real frames, ρ is the Euclidean distance between the two centroids, α and v are the aspect ratio evaluation parameters and the balance factor, respectively, and c denotes the diagonal distance of the minimum closure region containing the predicted and real frames. b and b^{gt} denote the coordinates of the centroids of the predicted and real frames, w and w^{gt} are the widths of the predicted and real frames, respectively, h and h^{gt} are the heights of the predicted and real frames, respectively.

However, the CIOU only reflects the difference between aspect ratios but fails to accurately capture the width, height and actual difference between them and the confidence level. To solve this problem, the CIOU is replaced using the EIOU loss function, which adds loss calculations for the width and height of the bounding box to the CIOU, taking into account not only the distance from the center point of the bounding box and the difference in area but also the difference in the width and height of the bounding box. Using the EIOU loss function can better measure the accuracy of predicting the location and shape of the bounding box and obtain more stable results in the target box regression process; hence, the model's prediction of the target bounding box will be more accurate and consistent in the training process. The EIOU loss function consists of three parts, which are the center distance loss L_{dis}, the overlap loss L_{IOU} and the width and height loss L_{asp} between the prediction frame and the real frame. The formula for the EIOU loss function is shown in Equation (5).

$$\begin{aligned} L_{EIOU} &= L_{IOU} + L_{dis} + L_{asp} \\ &= 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{C_w^2} + \frac{\rho^2(h, h^{gt})}{C_h^2} \end{aligned} \tag{5}$$

where C_w and C_h are the width and height of the two borders covered.

3. Binocular Vision Positioning Technology

Binocular vision is a kind of stereoscopic vision that simulates real vision to explain complex three-dimensional structures and spatial relationships and can provide depth and spatial information [23,24]. Binocular vision technology can help human beings to complete some detection and recognition tasks more accurately, conveniently and in real time when their eyes cannot judge accurately. Its basic principle is to use two eyes, according to two different images, after rectification, to match the pixels between the left and right images to form parallax, which is used to restore the depth and position information of the target in real three-dimensional space.

Binocular Positioning Principle

The binocular vision system compares the images obtained by the left and right cameras, obtains spatial, three-dimensional information by analyzing and calculating the parallax and calculates the distance between the object and the camera. Parallax calculation is based on image content. The more edge points, the higher the accuracy of the parallax calculation [25]. In addition, the accuracy of parallax calculations can also be improved by using color and shadow information. By calculating the parallax, the binocular vision system can recover the depth of the object. In computer vision, this depth information can be used to construct the three-dimensional structure of the scene and detect objects.

In an ideal situation, the binocular camera obtains the target picture using two identical cameras arranged in parallel, where each pixel row of the two cameras corresponds to each other one by one. The target distance information is then obtained through correction matching and calculation, as shown in Figure 5, which is the schematic diagram of binocular ranging.

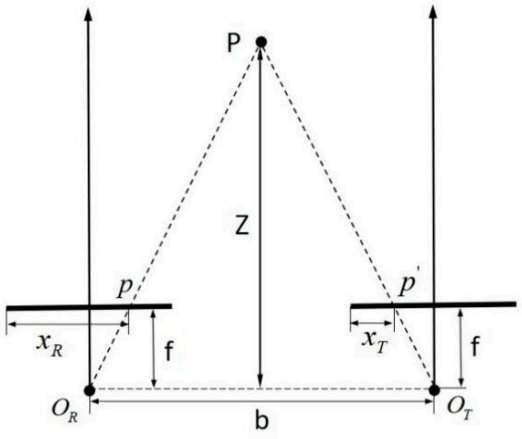

Figure 5. Schematic diagram of binocular positioning.

In the figure, f is the focal length of the camera; point P is the target point in the physical world; b is the baseline; O_R and O_T are the optical centers of the left and right cameras; p and p' are the imaging points of point P in the left and right camera image coordinate systems, respectively; Z is the distance between point P and the camera; x_R and x_T are the abscissas of the projection of point P in space in the images of the left and right cameras, respectively.

According to the principle of binocular ranging, in an ideal state, the parallax of the image we obtain is inversely proportional to the depth information (distance), with the parallax of the image represented by d [26]. According to the principle of triangle similarity, we can obtain the relationship shown in Formulas (7) and (8) and calculate the value of Z.

$$d = x_R - x_T \tag{6}$$

$$\frac{b - (x_R - x_T)}{Z - f} = \frac{b}{Z} \tag{7}$$

$$Z = \frac{b \cdot f}{x_R - x_T} \tag{8}$$

where d is the parallax of the left and right cameras, and the focal length and baseline of the camera can be obtained by camera calibration.

4. Results

4.1. Calibration of Binocular Camera

In binocular vision, the calibration process of the binocular camera actually determines the relationship between the position of each point in the physical world and the position of the pixel in the two-dimensional image [27]. According to this conversion process, the conversion relationship of any object in space from the world coordinate system to the pixel coordinate system can be obtained, as shown in Formula (9).

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{f}{dx} & 0 & u_0 & 0 \\ 0 & \frac{f}{dy} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = M_1 M_2 \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{9}$$

where (u, v) are the coordinates of P in the pixel coordinate system; and the coordinates of point P in the world coordinate system are, (X_w, Y_w, Z_w) where R is the rotation matrix and T is the translation matrix. M_1 is the internal reference matrix of the binocular camera, and M_2 is the external reference matrix of the binocular camera. The binocular camera takes pictures of the checkerboard grid calibration plate at different angles, as shown in Figure 6.

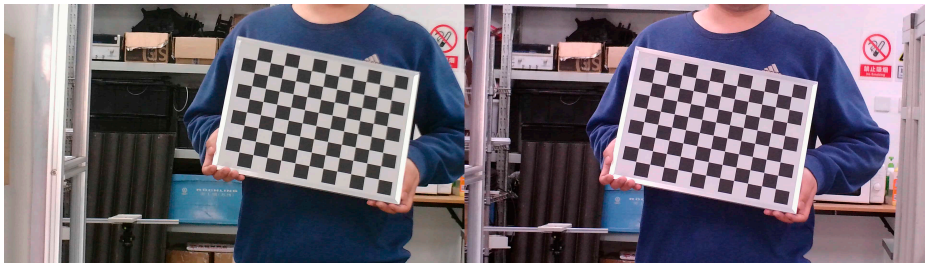

Figure 6. Pictures of collection calibration plates.

Using the calibration toolbox in Matlab2016b to process the photographic data taken, information about the characteristic corner points in the image is obtained by detecting the corner points in each picture of the image; the calibration of the camera is completed by detecting the corner points in different images, extracting the tessellated corner points and estimating the aberration error and other operations. Then, the relevant parameter information of the binocular camera is obtained, as shown in Table 1.

Table 1. Binocular camera calibration parameters.

Parameters	Left Camera	Right Camera
Internal reference matrix	$\begin{bmatrix} 1065.4700 & 0 & 961.2400 \\ 0 & 1065.9399 & 571.6750 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1066.1200 & 0 & 947.2400 \\ 0 & 1066.8199 & 573.0320 \\ 0 & 0 & 1 \end{bmatrix}$
Radial distortion	$\begin{bmatrix} -0.0488 & 0.0202 & -0.0083 \end{bmatrix}$	$\begin{bmatrix} -0.0517 & 0.0239 & -0.0083 \end{bmatrix}$
Tangential distortion	$\begin{bmatrix} 0.0002 & -0.0005 \end{bmatrix}$	$\begin{bmatrix} 0.0002 & -0.0005 \end{bmatrix}$
Rotation matrix	$\begin{bmatrix} 1.0000 & -0.0015 & -0.0030 \\ 0.0015 & 1{\cdot}0000 & 0.0003 \\ 0.0030 & -0.0003 & 1.0000 \end{bmatrix}$	
Translation matrix	$\begin{bmatrix} -119.841 & -0.3186 & 0.9505 \end{bmatrix}$	

4.2. Binocular Stereo Matching

Binocular stereo matching is the process of calculating the depth information of a 3D scene from images taken from two viewpoints. Among them, the parallax, which is the horizontal mean offset of pixels at the same position in the left and right images, is the key factor for calculating the depth [28]. Considering the real-time nature of the system application as well as the accuracy issue, the Semi-Global Matching (SGBM) algorithm, which is better in terms of accuracy and speed, is chosen [29]. The SGBM algorithm inherits the advantages of both Global Matching and Dynamic Programming and makes use of the optimal energy function to minimize the global function of the whole image by searching for the optimal parallax at each pixel point. An SAD window is used in the matching process to find the matching region of the left and right images to find the parallax. The calculation of the SAD cost is shown in Equation (10).

$$C(x,y,d) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} |L(x+i,y+j) - R(x+d+i,y+j)| \tag{10}$$

After the parallax map is obtained by the semi-local matching algorithm, the depth information of a point x on the image can be calculated based on the relevant information. The matching process is schematically shown in Figure 7, which can be used to obtain the value of parallax d. In this way, the depth information of the whole image can be calculated.

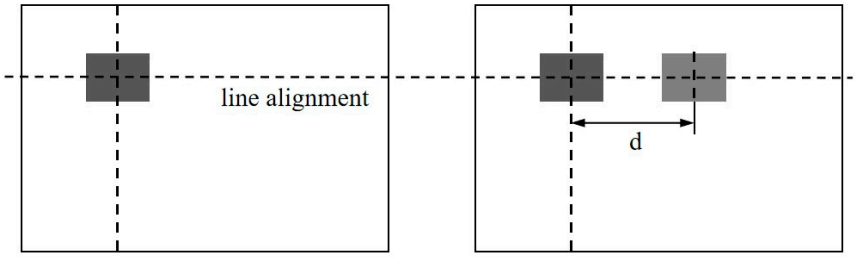

Figure 7. Stereo matching schematic.

5. Design of Vehicle Detection and Positioning System and Analysis of Experimental Results

5.1. Design of Vehicle Detection and Positioning System

System Software Design

The hardware of the vehicle detection and ranging positioning system designed in this research mainly includes the following components: An NVIDIA's Jetson NX embedded development board, serving as the computational core, with high computational performance and low-power characteristics. It features 384 CUDA cores, providing excellent parallel computing ability, and is equipped with a six-core ARM Cortex-A57 CPU, offering fast and reliable computing ability. The board is used for edge computing and AI algorithm deployment and offers high-performance GPUs and CPUs, rich scalability and excellent energy consumption control. Jetson NX, as the core hardware component of the system, provides powerful computational power and functional support for realizing vehicle detection, ranging and positioning.

In order to obtain high-quality binocular images, a binocular camera from HubVision is selected. The camera is characterized by high resolution, a high frame rate and low noise, which can provide clear and accurate images for depth information acquisition. The camera is connected to the Jetson NX through a USB3.0 interface.

The vehicle detection and localization system is based on the Jetson NX embedded development board. It is equipped with a binocular camera and a portable display, and it uses the Linux operating system for environment building and development. With the improved target detection algorithm and the fused binocular vision localization function, the system is able to achieve accurate vehicle detection, ranging and localization and has the potential to be applied in a variety of scenarios. The hardware composition of the system is shown in Figure 8.

Figure 8. System hardware design.

During the construction of the system, we chose to use algorithms deployed for vehicle detection and localization on an embedded development board, the Jetson NX. We improved the YOLOv5 target detection algorithm to improve detection capability and accuracy. We also realized the recovery of the spatial position of the detected vehicle by fusing the binocular vision localization function with the principle of stereo vision. The system software mainly includes four parts: binocular camera calibration, binocular vision stereo matching, vehicle detection with the improved YOLOv5 algorithm and vehicle ranging and localization.

Binocular camera calibration determines the geometric relationship between binocular cameras by acquiring a set of images, extracting feature points and using a calibration algorithm to calculate the internal and external parameters of the cameras to establish a mapping relationship between the left and right cameras [30]. Binocular vision stereo matching realizes the inference of depth information by comparing the images of the left and right cameras to find the corresponding feature points or pixels. By cCombining stereo matching with camera calibration, accurate 3D spatial information can be obtained. The improved YOLOv5 algorithm realizes real-time and accurate vehicle detection, utilizing a convolutional neural network and improving upon the framework of YOLOv5 to improve the accuracy and robustness of vehicle detection.

Based on the depth information of the binocular vision system and the target detection results, the distance and position of the vehicle can be calculated. The parallax value obtained by stereo matching is combined with the internal and external parameters of the camera to calculate the 3D coordinates relative to the camera using the principle of triangulation [31]. By aligning the camera coordinate system with the real-world coordinate system, the distance and position of the vehicle in the real world can be obtained. The software design flow chart is shown in Figure 9.

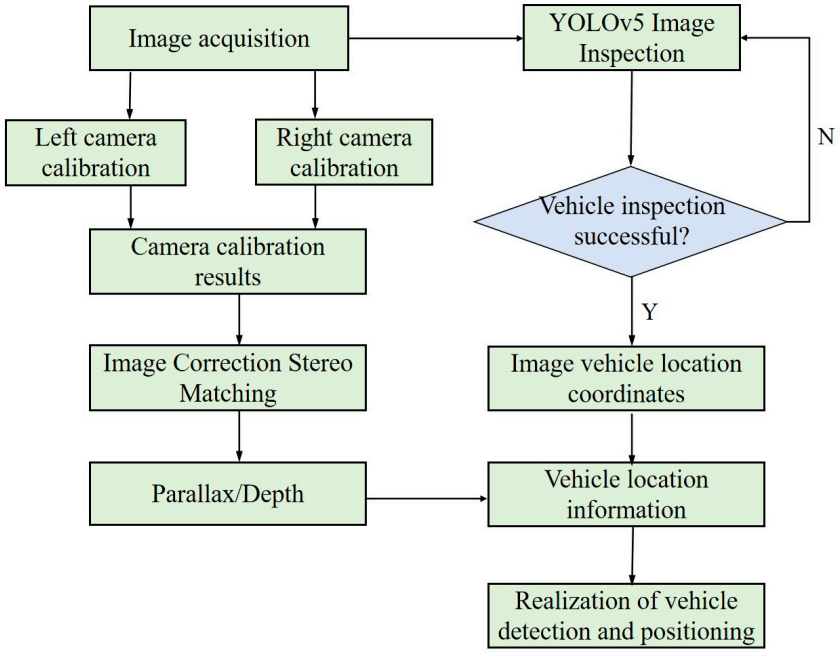

Figure 9. System software design flowchart.

5.2. Experimental Dataset

This experiment uses a dataset fused with some web images as well as vehicle images taken and collected by us. The dataset contains a variety of scenarios and can be used to simulate different environments in which vehicles are located. The dataset consists of 1760 images with a resolution of 1280 × 720, which are divided into training, testing and validation sets in a ratio of 7:2:1. We used a labeling tool to annotate these images and enhanced the diversity of the additional augmented training data in the image pre-processing

stage to improve the generalization ability of the detection model. Such processing greatly enriches the range of detection backgrounds and targets. The effect of data enhancement is shown in Figure 10.

Figure 10. Data enhancement display.

5.3. Vehicle Inspection Experiment

In order to verify the feasibility, real-time performance and accuracy of the improved YOLOv5 algorithm combined with binocular ranging for vehicle detection and localization, as well as to consider the convenience of the system, the algorithm is implemented under the Linux operating system on the embedded device. The PyTorch architecture is chosen to conduct experimental detection on the real captured images.

In this paper, recall, precision, mean average precision (mAP) and average accuracy (AP) are used to evaluate the performance of the improved model. Recall measures the model's ability to recognize positive examples, and precision measures the model's accuracy [32]. Mean average accuracy (mAP) is the average of the accuracies for different categories, which can be used to comprehensively assess the overall performance of the model. These metrics provide a comprehensive assessment of the performance of the improved model in terms of accuracy and precision.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$AP = \int_0^1 P(R)dR \tag{13}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{14}$$

where TP is the number of positive samples predicted correctly, FN is the number of negative samples predicted incorrectly, FP is the number of positive samples predicted

incorrectly and *TN* is the number of negative samples predicted correctly [33,34]. AP_i is the detection accuracy of category i, and *N* is the number of categories.

In order to evaluate the effect of the improved YOLOv5 target detection, three target detection models (YOLOv3, YOLOv5 and the improved YOLOv5) are chosen for comparative analysis. We can more accurately understand the improved algorithm in terms of performance enhancement, and the results are shown in Table 2.

Table 2. Algorithm model comparison results.

Algorithm	Precision (%)	Recall (%)	mAP.5 (%)
YOLOv3	81.9	71.9	85.4
YOLOv5	83.2	72.8	86.7
Improved YOLOv5	86.7	74.6	89.7

The detection of vehicles using the improved YOLOv5 target detection algorithm involves increasing the attention mechanism used in the vehicle detection task. By automatically learning the key vehicle feature regions, the attention mechanism reduces attention to irrelevant regions to improve the accuracy of detection. Using the model trained before and after the improvement to detect vehicles in the same scene, the improved model has a better effect on the detection of occluded regions as well as small target vehicles. The effect of vehicle detection before and after the improvement is shown in Figure 11.

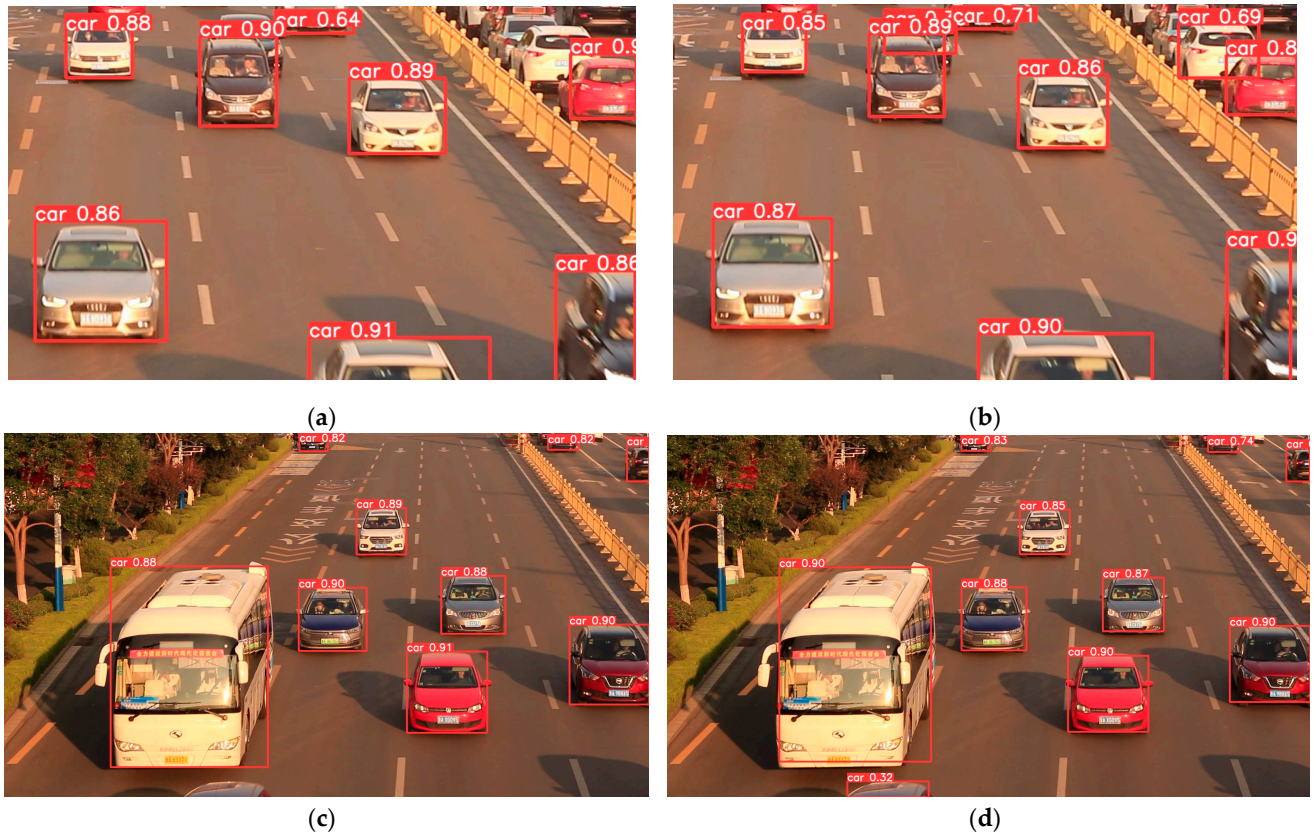

Figure 11. Vehicle detection and localization effect diagram. (**a**,**c**) The effect of unimproved detection; (**b**,**d**) the effect of improved detection.

By comparing (a) and (b) in the figure, we find that the improved algorithm performs better in vehicle detection. The number of vehicles detected by the pre-improved algorithm is 8, while it increases to 10 after the improvement. By comparing (c) and (d) in the figure, we can clearly see that the improved algorithm also has obvious advantages in the detection

scenarios. The improved algorithm not only recognizes the same number of vehicles as before the improvement but also captures more incomplete vehicle tail information.

Overall, by improving the loss function and introducing the attention mechanism, our algorithm has made significant progress in vehicle detection and target recognition. Especially in the occlusion region and small target region, the improved algorithm is able to better recognize and capture vehicle information.

5.4. Vehicle Ranging and Positioning Experiment

The experimental process using the improved vehicle detection model combined with binocular vision positioning technology involved vehicle detection and positioning experiments. In order to verify the effect of vehicle detection and positioning, the experiments were conducted on vehicles without using positional distance. When combined with binocular ranging, accurate detection of the vehicle was achieved. In addition, distance information relative to the binocular camera was shown at the top of the frame, enabling accurate implementation of the detection and positioning function. From the positioning effect shown in Figure 12, it can be seen that the vehicle detection and positioning system proposed in this paper can accurately detect the vehicle information in front and output the corresponding distance. This method enables real-time and accurate completion of the detection and positioning of the vehicle.

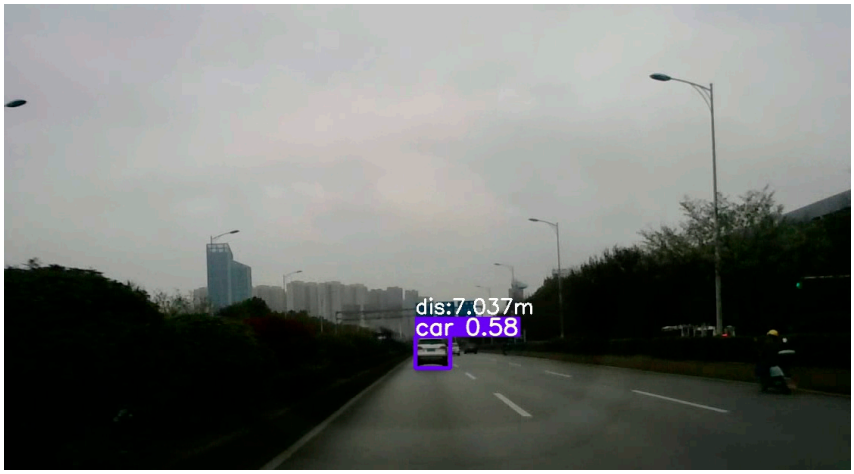

Figure 12. Vehicle positioning effect.

In order to ensure the effectiveness of localization and address localization accuracy issues, especially for different distances under the target vehicle, a number of detection and localization experiments were carried out. The measured distances were compared with the actual distances, and the results are shown in Table 3.

Table 3. Ranging and positioning results.

Serial Number	Actual Value (m)	Measured Value (m)	Relative Error (%)
1	2.000	2.011	0.550%
2	4.000	3.975	0.625%
3	6.000	5.957	0.717%
4	8.000	8.067	0.838%
5	10.000	10.114	1.140%
6	12.000	11.829	1.425%
7	14.000	14.261	1.864%
8	16.000	16.327	2.044%

From Table 3, it can be seen that in the range of 16 m, the ranging and positioning error of the vehicle is within 2.05%, indicating a better measurement and positioning effect. In

the actual positioning application process, there is high positioning accuracy, especially in the process of close-range use, where the positioning error is maintained at the centimeter level. In the process of long-distance positioning, the error will increase slightly, but the system will still be able to maintain positioning accuracy to ensure that the data are valid for the vehicle detection and positioning tasks, meeting the needs of the task.

6. Conclusions

In this study, we explored the binocular vision-based YOLOv5 vehicle detection and localization system and performed the improvement and integration of related algorithms. By introducing the attention mechanism and improving the loss function, we successfully enhanced the performance of the YOLOv5 algorithm in the vehicle detection task. Meanwhile, by integrating binocular vision information into the system, we could accurately estimate the position of the vehicle in space and thus locate the vehicle more accurately. The system could detect and localize vehicles in road scenes more accurately and handle vehicle detection and localization tasks in different road scenes, which can be used as a reference for vehicle perception work in automatic driving tasks and has certain application value.

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

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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

References

1. Zhang, X.; Gao, H.; Zhao, J. Overview of deep learning intelligent driving methods. *J. Tsinghua Univ. Sci. Technol.* **2018**, *58*, 438–444. [CrossRef]
2. Zhang, S.; Wang, Y.; Xiao, H. Pedestrian and Vehicle Detection Algorithm Based on Improved YOLOv5 in Haze Weather. *Rad. Eng.* **2023**, 1–10. Available online: http://kns.cnki.net/kcms/detail/13.1097.TN.20230726.1847.012.html (accessed on 27 July 2023).
3. Hu, J.; Wang, H.; Dai, X. Real-Time Detection Algorithm for Small-Target Traffic Signs Based on lmproved YOLOv5. *Comput. Eng. Appl.* **2023**, *59*, 185–193.
4. Han, Z.; Wang, H.; Wu, X. Monocular vision detection and localization method of UAV based on YOLOv5. *Flig. Dynam.* **2023**, *41*, 61–66+81. [CrossRef]
5. Tang, T.; Zhou, S.; Deng, Z. Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining. *Sensors* **2017**, *17*, 336. [CrossRef] [PubMed]
6. Ren, S.; He, K.; Ross, G.; 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]
7. Jia, Z.; Wang, W.; Liu, G. Improved Lightweight Traffic Sign Detection Algorithm of YOLOv5. *J. Data Acquis. Process.* **2023**, *38*, 1434–1444. [CrossRef]
8. He, M.; Zhang, X.; Zhang, X. Research on plane workpiece recognition and localization based on binocular stereo vision. *J. Laser.* **2023**, *44*, 199–204. [CrossRef]
9. Zhang, X.; Zhao, J.; Wang, S. Binocular vision detection and positioning method for ships based on YOLOv3 algorithm. *J. Shanghai Mari. Univ.* **2021**, *42*, 26–32. [CrossRef]
10. Chen, Z.; Wang, X.; Zhang, W.; Yao, G.; Li, D.; Zeng, L. Autonomous Parking Space Detection for Electric Vehicles Based on Improved YOLOV5-OBB Algorithm. *World Electr. Veh. J.* **2023**, *14*, 276. [CrossRef]
11. Niu, C.; Song, Y.; Zhao, X. SE-Lightweight YOLO: Higher Accuracy in YOLO Detection for Vehicle Inspection. *Appl. Sci.* **2023**, *13*, 13052. [CrossRef]
12. Shao, L.; Wu, H.; Li, C.; Li, J. A Vehicle Recognition Model Based on Improved YOLOv5. *Electronics* **2023**, *12*, 1323. [CrossRef]
13. Hu, P.; Wang, Y.; Zhai, Q. Research on Night Vehicle Detection Algorithm Based on YOLOv5s and Bistable Stochastic Resonance. *Comput. Sci.* **2024**, 1–11. Available online: http://kns.cnki.net/kcms/detail/50.1075.TP.20231120.1033.022.html (accessed on 20 November 2023).
14. Yao, J.; Fan, X.; Li, B.; Qin, W. Adverse Weather Target Detection Algorithm Based on Adaptive Color Levels and Improved YOLOv5. *Sensors* **2022**, *22*, 8577. [CrossRef] [PubMed]

15. Zhang, C.; Guo, C.; Li, Y. Research on Aircraft Door Identification and Position Method Based on Improved YOLOv5. *Comput. Meas. Cont.* **2024**, 1–9. Available online: http://kns.cnki.net/kcms/detail/11.4762.TP.20230816.1145.022.html (accessed on 16 August 2023).
16. Pan, X.; Jia, N.; Mu, Y. Improved YOLOv4-Based Small Object Detection Method in Complex Scenes. *Int. J. Pattern Recognit. Artif. Intell.* **2023**, *37*, 2350024. [CrossRef]
17. Yang, L.; Liu, S.; Zhao, Y. Deep-Learning Based Algorithm for Detecting Targets in Infrared Images. *Appl. Sci.* **2022**, *12*, 3322. [CrossRef]
18. Wu, T.; Wang, T.; Liu, Y. Real-Time Vehicle and Distance Detection Based on Improved Yolo v5 Network. In Proceedings of the 2021 3rd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 18–20 June 2021; pp. 24–28. [CrossRef]
19. Yuan, P.; Cai, D.; Cao, W. Train Target Recognition and Ranging Technology Based on Binocular Stereoscopic Vision. *J. Northeast. Univ. Nat. Sci.* **2022**, *43*, 335–343.
20. Chen, S.; Lin, W. Embedded System Real-Time Vehicle Detection based on Improved YOLO Network. In Proceedings of the 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 11–12 October 2019; pp. 1400–1403. [CrossRef]
21. Yao, A.; Xu, J. Electric car charging hole identification and positioning system based on binocular vision. *Trans. Micr. Technol.* **2021**, *40*, 81–84. [CrossRef]
22. Zaarane, A.; Slimani, I.; Okaishi, A. Distance measurement system for autonomous vehicles using stereo camera. *Array* **2020**, *5*, 100016. [CrossRef]
23. Wang, L.; Duan, J.; Xin, L. YOLOv5 Helmet Wear Detection Method with Introduction of Attention Mechanism. *Comput. Eng. Appl.* **2022**, *58*, 303–312.
24. Amrouche, A.; Bentrcia, Y.; Abed, A.; Hezil, N. Vehicle Detection and Tracking in Real-time using YOLOv4-tiny. In Proceedings of the 2022 7th International Conference on Image and Signal Processing and Their Applications (ISPA), Mostaganem, Algeria, 8–9 May 2022; pp. 1–5. [CrossRef]
25. Lv, H.; Lu, H. Research on traffic sign recognition technology based on YOLOv5 algorithm. *J. Electron. Meas. Instrum.* **2021**, *35*, 137–144. [CrossRef]
26. Luo, G.; Chen, X.; Lin, W.; Dai, J.; Liang, P.; Zhang, C. An Obstacle Detection Algorithm Suitable for Complex Traffic Environment. *World Electr. Veh. J.* **2022**, *13*, 69. [CrossRef]
27. Yang, R.; Yu, S.; Yao, Q.; Huang, J.; Ya, F. Vehicle Distance Measurement Method of Two-Way Two-Lane Roads Based on Monocular Vision. *Appl. Sci.* **2023**, *13*, 3468. [CrossRef]
28. Zhang, H.; Zhang, T.; Ren, Y. A Fast Binocular Vision Stereo Matching Algorithm. *Appl. Mech. Mater.* **2014**, *3027*, 3735–3738. [CrossRef]
29. Zhao, W.; Wu, S.; Zhang, Y. Multi-target obstacle tracking and ranging based on deep learning and binocular vision. *J. Laser.* **2023**, *44*, 57–64. [CrossRef]
30. Xu, H.; Wang, L.; Chen, F. Advancements in Electric Vehicle PCB Inspection: Application of Multi-Scale CBAM, Partial Convolution, and NWD Loss in YOLOv5. *World Electr. Veh. J.* **2024**, *15*, 15. [CrossRef]
31. Zhang, P.; Liu, J.; Xiao, J. Target localization and tracking method based on camera and lidar fusion. *Laser Optoel. Progr.* **2023**, 1–16. Available online: http://kns.cnki.net/kcms/detail/31.1690.TN.20230821.1446.134.html (accessed on 22 August 2023).
32. Tong, Z.; Zhao, T.; He, L. Localization and Driving Speed Detection for Construction Vehicles Based on Binocular Vision. *China Mech. Eng.* **2018**, *29*, 423–428.
33. Wen, X.; Xiao, H.; Wang, D.; Cao, X. Research on Marine Hoisting Location Based on Binocular Vision and ToF. *Instr. Technol. Sens.* **2023**, *6*, 121–126.
34. Wei, C.; Yang, R.; Liu, Z. YOLOv8 with bi-level routing attention for road scene object detection. *J. Graph.* **2023**, *44*, 1104–1111. [CrossRef]