applied sciences

MDPI

Article

Binocular Vision-Based Pole-Shaped Obstacle Detection and Ranging Study

Lei Cai [1,2], Congling Zhou [1,2,*], Yongqiang Wang [1,2], Hao Wang [1,2] and Boyu Liu [1,2]

[1] School of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300222, China; cailei@mail.tust.edu.cn (L.C.); wangyq@tust.edu.cn (Y.W.); wangtn199011@tust.edu.cn (H.W.); liuby@mail.tust.edu.cn (B.L.)

[2] Tianjin Key Laboratory for Integrated Design & Online Monitor Center of Light Design and Food Engineering Machinery Equipment, Tianjin University of Science &Technology, Tianjin 300222, China

* Correspondence: zhoucling@tust.edu.cn

Abstract: (1) Background: In real road scenarios, various complex environmental conditions may occur, including bright lights, nighttime, rain, and snow. In such a complex environment for detecting pole-shaped obstacles, it is easy to lose the feature information. A high rate of leakage detection, false positives, and measurement errors are generated as a result. (2) Methods: The first part of this paper utilizes the improved YOLOv5 algorithm to detect and classify pole-shaped obstacles. Then, the identified target frame information is combined with binocular stereo matching to obtain more accurate distance information. (3) Results: The experimental results demonstrate that this method achieves a mean average precision (mAP) of 97.4% for detecting pole-shaped obstacles, which is 3.1% higher than the original model. The image inference time is only 1.6 ms, which is 1.8 ms faster than the original algorithm. Additionally, the model size is only 19.0 MB. Furthermore, the range error of this system is less than 7% within the range of 3–15 m. (4) Conclusions: Therefore, the algorithm not only achieves real-time and accurate identification and classification but also ensures precise measurement within a specific range. Meanwhile, the model is lightweight and better suited for deploying sensing systems.

Keywords: complex environment; binocular stereo vision; object detection; YOLOv5; real-time and accuracy

Citation: Cai, L.; Zhou, C.; Wang, Y.; Wang, H.; Liu, B. Binocular Vision-Based Pole-Shaped Obstacle Detection and Ranging Study. *Appl. Sci.* **2023**, *13*, 12617. https://doi.org/10.3390/app132312617

Academic Editors: Lifei Wei, Liqin Cao and Xuan Zhang

Received: 13 October 2023
Revised: 17 November 2023
Accepted: 20 November 2023
Published: 23 November 2023

1. Introduction

Traffic safety [1] has always been the focus of the world's attention and one of the main problems solved by each country. In recent years, research institutes, automobile companies, and other industries have been conducting in-depth research on intelligent driving, which is typically represented by Advanced Driver Assistance Systems (ADAS) [2,3], which can monitor the road and traffic environment, as well as the state of the vehicle itself, and then help drivers prevent accidents and improve driving safety through warnings, interventions, or autonomy control. The first task for ADAS is obstacle avoidance, but the prerequisite for obstacle avoidance is the accurate detection of obstacles. At present, although there are many different methods for target detection, there are various shortcomings, such as: missed detection, false detection, lack of real time, and so on. According to the research, it is found that the current environment perception of ADAS systems mainly has the application based on a radar point cloud [4] and the application based on pure vision [5]. Although radar technology has a long measuring distance, the difficulties of not being able to obtain visual information, the high cost, and the difficulty in distinguishing between multiple targets have been the difficulties of research in various industries. For visual perception, although limited by the complex environment, a high resolution can provide rich target information and is favored in the car industry due to its low cost and other advantages. The typical representative among them is the 2023 Tesla FSD V12, which is an end-to-end

pure visual perception system. The detection of obstacles is one of the key research focuses of the environment perception system, and the obstacle detection methods are mainly divided into traditional visual detection and deep learning-based target detection. With the rapid development of computer vision technology and deep learning algorithms, obstacle detection has become an important technology in assisted driving systems [6].

Currently, deep learning-based obstacle detection methods have become mainstream, and the commonly used architectures are R-CNN [7], Fast R-CNN [8], Faster R-CNN [9], mask R-CNN [10], and YOLO [11]. Among them, the YOLO series of algorithms performs better in terms of speed and accuracy and is one of the most widely used obstacle detection algorithms.

However, in practical applications, pole-shaped obstacles are susceptible to environmental factors such as lighting, weather conditions, and occlusion. These factors can significantly impact the accuracy of detection and the range of pole-shaped obstacles. Moreover, due to the different heights and shapes of pole-shaped obstacles, multi-scale detection is required to better detect pole-shaped obstacles of different sizes and shapes. Dhall et al. [12] proposed a method for the rapid detection of traffic cones. First, an improved detector is used to detect the traffic cones. Then, a regression network is employed to identify the key features of the traffic cones. Finally, the three-dimensional information of the traffic cones is obtained using the perspective n-point algorithm. Although this method can run on low-power hardware, there is still significant room for improvement in terms of detection speed and accuracy. He et al. [13] propose a feature fusion method aimed at improving obstacle detection performance under foggy conditions. The main idea is to identify the differences in image features between sunny and foggy days, construct a dataset of foggy images, and then utilize the feature fusion method to enhance obstacle detection performance. Liu et al. [14] proposed a method of fusing convolutional features and the GCANet network for the problem of the difficult extraction of obstacle feature information under a foggy sky. After processing the obstacle images of the pair under foggy weather, the information on the original obstacles is fully retained, thus realizing the detection of obstacles under the foggy sky. Pan et al. [15] proposed a method for integrating YOLO and monocular vision techniques for the detection of pedestrian distance in complex environments. To realize the detection of distance between multiple pedestrians in complex environments, Luo et al. [16] proposed a stereo vision-based method for roadless spatial extraction and obstacle detection for the task of obstacle detection in complex transportation environments. The method was based on a V parallax image and RANSAC algorithm, which enables obstacle detection by extracting the height and width information of obstacles on the road. Guan et al. [17] proposed a method that fuses YOLOv4 and binocular stereo vision to reduce the cost of autonomous driving environment perception methods, among other issues. Although many researchers have devoted themselves to the study of real-time and accurate target detection, they have not reached a good balance of the detection accuracy and detection rate.

In particular, this problem is more prominent in the research on target ranging and recognition. Therefore, it is of great significance for the development of perception technology in assisted driving systems to quickly and accurately obtain the key features of pole-shaped obstacles in complex and changing environments. The contributions of this study are as follows.

1. The CIoU loss function is replaced by the SIoU loss function, which is improved to address the issue of overlapping multiple prediction frames and the optimal matching of anchor frames, while also ensuring real-time performance. Meanwhile, the improved loss function is redefined and named Monge–Kantorovich SIoU (MKS).
2. A multi-scale feature efficient fusion network architecture (MFFNA) is proposed. It extracts feature information from different scales of the feature space through an efficient multi-scale feature extraction module.
3. A hybrid attention mechanism is introduced. The feature information is passed to the hybrid attention mechanism to improve the extraction of multi-scale information

in the feature space, suppress irrelevant and complex environmental background information, and focus on the feature information of road obstacles.

4. The detected target frame information is fused with the binocular stereo-matching algorithm, enabling the accurate recognition, classification, and ranging of pole-shaped obstacles in complex environments.

The rest of the paper is organized as follows. Section 2 describes the core algorithms, including the improved target detection algorithm and the target detection algorithm fused with the binocular ranging algorithm. Section 3 describes the experimental environment, datasets, and evaluation metrics for the model. Section 4 analyzes the enhanced network model experimentally. Finally, conclusions are drawn in Section 5.

2. Methods

First, the images acquired by the binocular camera in real time are inputted to the ranging module with binocular images and to the detection module with left-eye images, respectively. Then, stereo correction and stereo matching are performed in the ranging module, respectively. At the same time, the target information is accurately obtained in the target detection module. Finally, the target information and stereo-matching are correlated, allowing for more accurate species and distance information to be obtained. The overall process is shown in Figure 1.

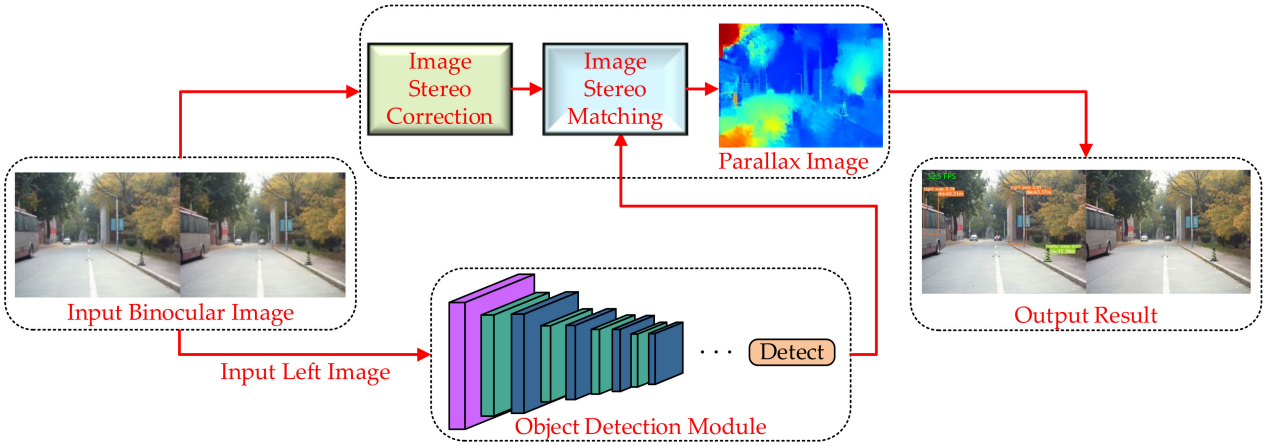


Figure 1. Overall workflow diagram.

2.1. MKS Loss Function

Object detection and recognition classification are key research areas in computer vision, and the loss function plays an important role in determining detection accuracy. The loss function is used to evaluate the difference between the model's prediction result and the actual target. The smaller the loss function value, the closer the prediction result is to the actual target. The loss function of YOLOv5 consists of three main components: rectangular box loss (box loss), confidence loss (obj loss), and classification loss (cls loss). Therefore, the total loss function of the YOLOv5 algorithm is defined as follows:

$$\text{Total loss} = A \times \text{obj loss} + B \times \text{cls loss} + C \times \text{box loss} \quad (1)$$

where, A, B, and C are the weight values of the three loss functions and A = 1, B = 0.5, and C = 0.1.

In the current study, YOLOv5 utilizes the CIoU loss function [18]. However, this loss function only considers the aggregation of the bounding box regression metrics and does not account for the mismatch between the required true and predicted frames. As a result, it leads to slow convergence and low efficiency. Therefore, the SIoU loss function [19] is used instead of the original loss function. However, there are some problems with the SIoU loss function, such as category imbalance in rod obstacles, the overlapping of multiple

prediction frames, and the issue of optimal matching. Therefore, based on the SIoU loss function, the Monge–Kantorovich (MK) [20] function is introduced to optimize the IoU loss function. Specifically, the MK algorithm considers two sets of points as input values. In other words, the predicted bounding box and the true bounding box are treated as sets of points. Then, the distance matrix between these points is computed, where each entry in the matrix represents the distance between the predicted bounding box and the true bounding box. Finally, the Sinkhorn–Knopp algorithm [21] is used to find the best match between the predicted and real bounding boxes, minimizing the total negative IoU loss. In addition, the *MK* function can regularize the model to prevent overfitting. To better understand the *MK* function, it is defined as follows. The mathematical definition of the Monge problem [22] is: Given two metric spaces X and Y and two probability measures $\mu \in P(X)$, $\nu \in P(Y)$, loss function c: $X \times Y \to R \cup \{+\infty\}$:

$$(MP) := \inf\left\{\int_X c(x, T(x))d\mu(x): T_{\#\mu=\nu}\right\} \tag{2}$$

where c (x, T(x)) denotes the loss of x to T(x), $T_{\#\mu=\nu}$ denotes the mapping between μ and ν, and T extrapolates the probability measure μ to ν.

Kantorovich generalizes the transmission mapping to the joint probability distribution, which has:

$$\prod(\mu, \nu) = \{\gamma \in P(X \times Y) | \pi_{X\#\gamma} = \mu, \pi_{Y\#\gamma} = \nu\} \tag{3}$$

The Kantorovich problem [23] is mathematically defined as two probability measures $\mu \in P(X)$ and $\nu \in P(Y)$ and a cost function c:$X \times Y \to R \cup \{+\infty\}$ for two metric spaces X, Y:

$$(KP) := \inf\left\{\int_{X\times Y} c(x, y)d\gamma(x, y) | \gamma \in \prod(\mu, \nu)\right\} \tag{4}$$

In Monge's problem, only one type of mapping can be targeted, either one-to-one or many-to-one. However, the Kantorovich problem can be solved by simplifying the problem in a one-to-many manner.

The SIoU loss function is introduced to redefine the existing loss functions for the current problem, which include angle loss, distance loss, shape loss, and IoU loss. The IoU loss is the most commonly used loss function for target detection. It represents the intersection over the union ratio of the true and predicted frames. However, while the IoU metric is used for computation, negative IoU is used for optimization during training. Negative IoU is defined as follows.

$$\text{negative IoU} = \frac{(MP)\inf}{(KP)\inf} - \text{IoU}(P, G) \tag{5}$$

where P denotes the prediction frame and G denotes the true frame. Thus, the final expression of the SIoU loss function is as follows.

$$\text{Loss}_{MKS} = \text{negative IoU}(1 - \text{IoU}) + \frac{\Delta + \Omega}{2} \tag{6}$$

where Δ represents the distance cost, while Ω represents the shape cost. IoU represents the ratio of the union and intersection between the ground truth box and the predicted box.

2.2. Hybrid Attention Mechanisms

The attention mechanism [24] is a technique used to simulate human attention and has been widely applied in the field of deep learning. Currently, there are two main types of attention mechanisms for processing feature maps: channel attention mechanisms and spatial attention mechanisms [25]. The model in this paper is mainly used in outdoor complex environments to better obtain accurate detection information. It is very necessary to suppress the complex background and pay attention to the important feature informa-

tion. However, the above two attention mechanisms can be used not only to selectively focus on specific regions or channels in the input data but also to better capture useful feature information. Consequently, integrating these two mechanisms can not only play their respective advantages to improve the model's performance but can also reduce the computational complexity and enhance the generalization ability.

Currently, there are two main methods for connecting the two attention mechanisms in deep learning: the cascade and parallel methods. The parallel attention mechanism has a relatively simple structure and may not fully exploit the correlation between different feature subspaces, resulting in insufficient feature representation. Therefore, more layers need to be stacked. The cascade attention mechanism, in contrast, can utilize the outputs of multiple attention modules in a cascade. This allows for the better capture of spatial and channel correlations and enhances the feature expression capability. In addition, the cascade structure can also perform attention fusion for feature maps of different scales to better adapt to objects of varying sizes. The cascade structure is illustrated in Figure 2.

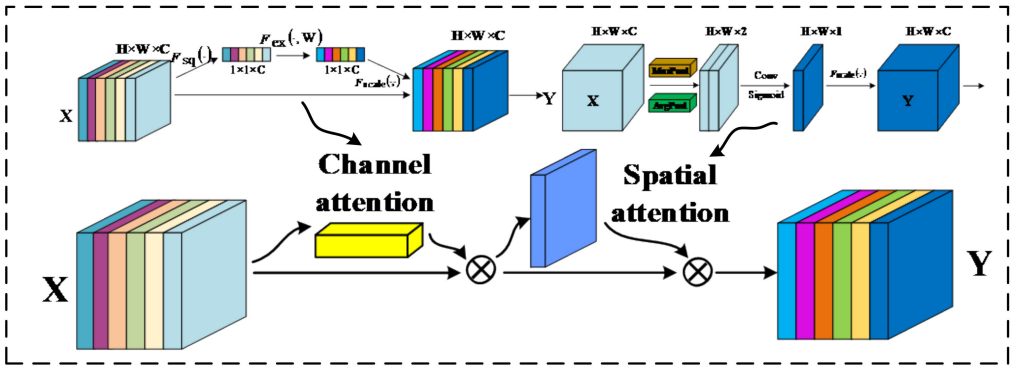


Figure 2. Cascade Hybrid Attention Mechanism.

2.3. Multi-Scale Feature Pyramid

The main types of objects detected in this paper include electric poles, surveillance poles, traffic signal poles, traffic cones, road pile poles, street lamp poles, and traffic sign poles. These pole-shaped obstacles vary in size, ranging from small traffic cones to large surveillance poles. Due to the significant variation in the scales of their features, this paper proposes an optimized multi-scale feature fusion network architecture (MFFNA).

The core idea is to introduce a multi-scale feature pyramid. However, traditional feature pyramid structures, such as FPN and PAN, have issues such as high computational costs and inadequate information propagation. Therefore, an efficient feature pyramid called Reap-GFPN [26] was introduced in this paper. Although Rep-GFPN addresses the issue of multi-scale features, it still has some limitations in terms of performance, including a high computational cost and large memory consumption.

First, this network model replaces the ConvBNAct and ConvWrapper operators with convolution (Conv) layers and Cross Stage Partial Network (CSPStage) to simplify the model structure, improve the training speed, and reduce the model complexity. Furthermore, the complexity of the model is reduced and the training speed is improved by replacing BepC3 with a lightweight convolution layer called C3. Finally, a hybrid attention mechanism called the Convolutional Block Attention Module (CBAM) [27] was introduced into the neck network of the Rep-GFPN to enhance attention towards important features and suppress the extraction of unimportant features. This improvement not only improves the accuracy and robustness of the model but also reduces the computational cost and memory consumption, and its network architecture is shown in Figure 3.

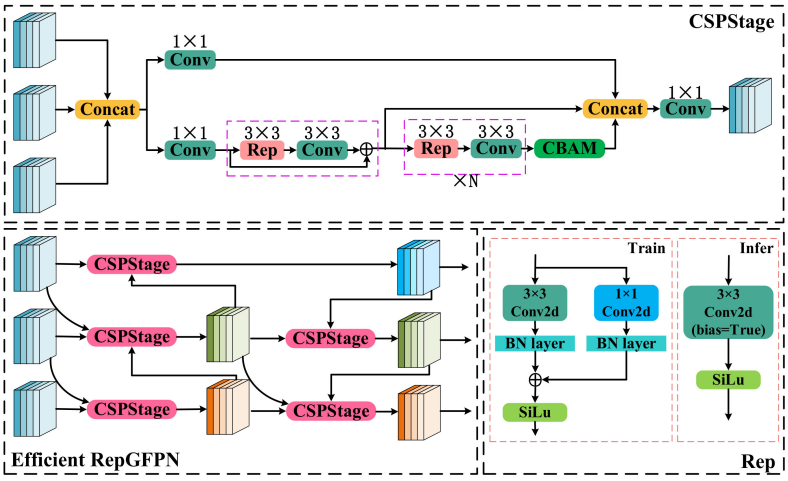


Figure 3. Multi-scale feature pyramid network structure.

According to the network architecture diagram, it can be seen that in the feature fusion module, this paper first utilizes the improved CSPStage [28] for convolutional fusion. Then, it proceeds with structural reparameterization—specifically, the fusion of Conv2d and BN. Therefore, the expression of BN in the channel i of the feature map is as follows.

$$y_i = \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \gamma_i + \beta_i \tag{7}$$

where μ denotes the mean, σ^2 denotes the variance, γ denotes the weights, β denotes the bias, and ε is a very small constant used to prevent the denominator in the formula from being zero.

2.4. Target Detection Network

Since pole-shaped obstacle detection systems for assisted driving require a certain level of accuracy and real-time performance, therefore, the improved MFMAM-YOLOv5s algorithm is used to train the model and perform target detection. The network structure of this algorithm is shown in Figure 4.

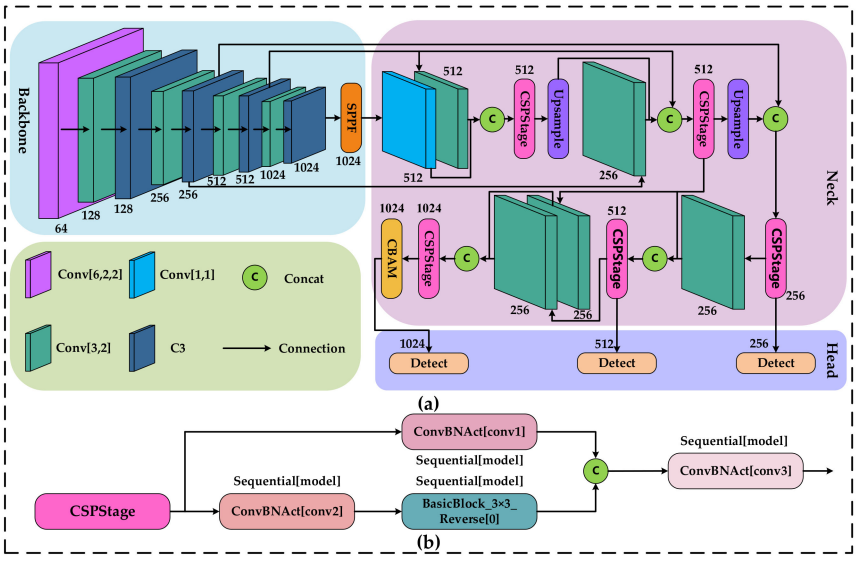


Figure 4. MFMAM-YOLOv5s network architecture diagram. (**a**) represents the main network diagram, while (**b**) illustrates the internal structure of CSPStage.

2.5. Binocular Camera Calibration

In this paper, the Zhang Zheng You calibration method [29] was chosen for binocular camera calibration. A total of 27 sets of images were collected, and the MATLAB calibration toolbox was used to extract the internal and external parameters of the camera, as well as the distortion coefficients. These calibration parameters were then utilized to perform stereo calibration on the collected left and right images. The results of the binocular camera calibration are presented in Table 1.

Table 1. Binocular camera calibration of internal and external parameters.

Parameters	Left Camera	Right Camera
Intrinsic Matrix	$\begin{bmatrix} 2186.8 & 0 & 655.5445 \\ 0 & 2188.3 & 512.99 \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 2114.0 & 0 & 647.918 \\ 0 & 2115.6 & 528.0488 \\ 0 & 0 & 1 \end{bmatrix}$
Distortion	$\begin{bmatrix} -0.3846 & 0.2401 & 0 & 0.003 & 0 \end{bmatrix}$	$\begin{bmatrix} -0.4261 & 0.1959 & 0 & 0.002 & 0 \end{bmatrix}$
Translation	$\begin{bmatrix} -119.6552 & -0.0624 & 0.6339 \end{bmatrix}$	
Rotating	$\begin{bmatrix} 0.9999 & -0.012 & -0.0116 \\ 0.0119 & 0.9999 & -0.0047 \\ 0.0117 & 0.0045 & 0.9999 \end{bmatrix}$	

2.6. Stereo Correction

Stereo correction [30] utilizes the internal and external parameters of binocular calibration, as well as the binocular relative position relationship, to eliminate aberrations and align the lines for the left and right views, respectively. In this paper, the Bouguet correction method [31] is used to minimize the number of re-projections for each of the left and right images, and the corrected image is shown in Figure 5.

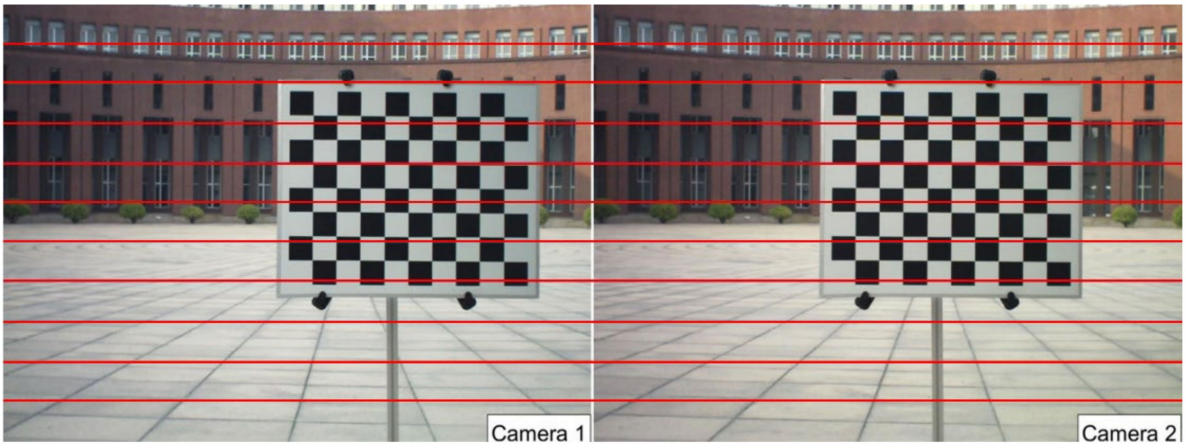


Figure 5. Stereo calibration chart.

2.7. Stereo Matching

Stereo-matching technology [32] is the most critical research topic in binocular ranging systems, and the accuracy of image matching directly affects the accuracy of subsequent binocular ranging. However, the stereo-matching process has some limitations and sources of error, such as disparity range limitation, illumination variation, and mismatching factors. These factors can lead to errors or inaccuracies in binocular stereo matching. To reduce these errors, this study first adopts object detection. Then, it associates the feature information obtained from object detection with binocular stereo matching, enriching the features and coordinate information. This enhances the accuracy of binocular ranging and reduces errors. Nevertheless, there are numerous existing stereo-matching algorithms. To meet the

requirements of accurate and fast stereo matching in complex scenes, this study selects the Semi-Global Block Matching algorithm (SGBM) [33] as the matching algorithm for bar-shaped obstacle features. At the same time, the SGBM stereo matching algorithm not only has high real-time performance and measurement accuracy but also possesses an efficient execution capability.

3. Experiments

3.1. Experimental Platform and Environment

To ensure the fairness and reliability of each experiment, all experiments in this chapter were conducted as shown in Table 2.

Table 2. Experimental environment.

Type	Parameter
GPU	NVIDIA GeForce RTX 3090 GPU
CPU	Xeon(R) Platinum 8255C CPU @ 2.50 GHz
Language version	Python 3.8.16
RAM/VRAM	45 GB/24 GB
Framework and gas pedal versions	Pytorch1.13.1, CUDA11.6, and cuDNN8.4.0

The hyperparameter settings in the experimental model are shown in Table 3.

Table 3. Hyperparameter settings.

Hyperparameter	Value
Epochs	300
Batch size	16
Optimizer	SGD
workers	8
Momentum	0.937

The above information pertains to the configuration of experimental hardware parameters and model hyperparameters. However, when performing target ranging, the ranging accuracy may vary due to the different specifications of the binocular camera. The specifications of the binocular camera used in this experiment are shown in Table 4.

Table 4. Parameters of the Binocular Camera.

Parameters	Value
Image resolution	1280 × 720
Shutter type	Global shutter
Size of the target surface	1/3 inch
Maximum frame rate	60 fps
Single pixel size	3.75 μm × 3.75 μm
Baseline	120 mm

3.2. Introduction to the Dataset

Currently, there is a significant lack of datasets for pole-shaped obstacles. Therefore, this paper created a static dataset called PSO 2023. The dataset collected pole-shaped obstacles in real scenes during all four seasons: spring, summer, autumn, and winter. The main idea is to capture images at different times of the day, including morning, noon, afternoon, and evening. To enhance the generality and robustness of the model, the dataset also includes 2800 images from complex environments, such as rainy or snowy weather, strong lighting, and obstructions, totaling 6300 images. These images were then divided into training, validation, and testing sets in an 8:1:1 ratio. As shown in Figure 6, this paper

mainly focuses on studying seven major categories, nine minor categories, and a hundred styles of pole-shaped obstacles in urban roads.

Figure 6. Categorization of detections. (**a**) Electric pole; (**b**) Surveillance pole; (**c**) Fixed traffic signal pole; (**d**) Mobile traffic signal pole; (**e**) Traffic cone; (**f**) Fixed road pile pole; (**g**) Mobile road pile pole; (**h**) Street lamp pole; (**i**) Traffic sign pole.

3.3. Data Augmentation

Data augmentation [34–36] is a commonly used technique in deep learning to increase the diversity of datasets and improve the generalization ability and robustness of models. In YOLOv5, data augmentation can be broadly classified into color transformations (e.g., noise, blur, and contrast) and geometric transformations (e.g., rotation, translation, and scaling). In this paper, the dataset was rotated and darkened, and Gaussian noise was added, respectively, as shown in Figure 7.

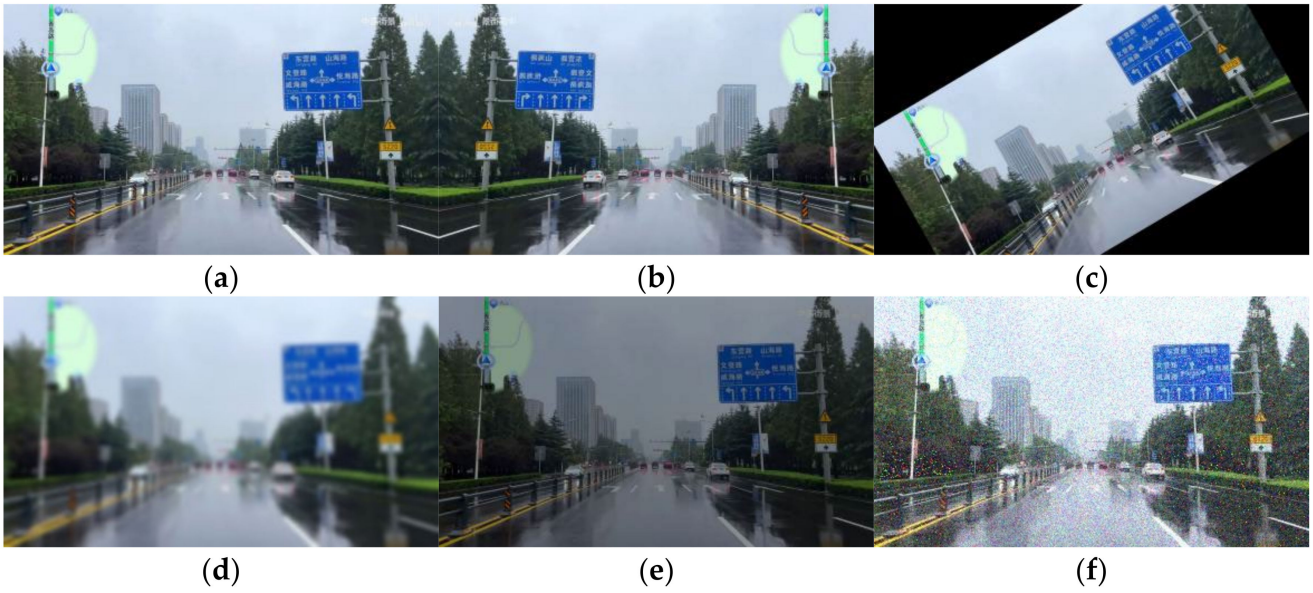


Figure 7. Data augmentation technology. (**a**) Original image; (**b**) Flip horizontal; (**c**) Random rotation; (**d**) Fuzzy processing; (**e**) Random brightness; (**f**) Gaussian noise.

3.4. Evaluation Metrics

The article mainly uses four evaluation indexes—precision, recall, mAP_0.5, and mAP_0.5:0.95—to judge the effectiveness of the model detection performance, and the formulas for each type of index are as follows.

$$P_{\text{precision}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$R_{\text{recall}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{AP} = \int_0^1 P(R)dR \tag{10}$$

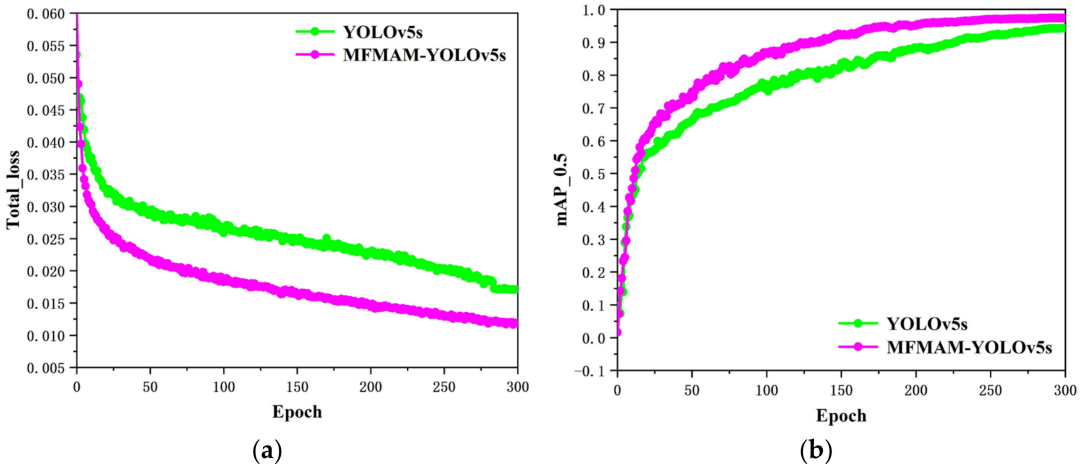
$$\text{mAP} = \frac{1}{\text{N}} \sum_{n \in \text{N}} \text{AP}(n) \tag{11}$$

where TP denotes the number of correctly detected targets, FP denotes the number of incorrectly detected targets, FN denotes the number of undetected targets, and N denotes the number of classes that need to be classified in total. Recall indicates the number of correctly predicted samples as a percentage of all samples that are positive cases. mAP can be used as a comprehensive evaluation metric for individual category detection. Higher AP values indicate better detection of a category, and mAP is a comprehensive evaluation of the entire network.

4. Analysis of Results

4.1. Model Training

In this paper, the original YOLOv5s algorithm and the MFMAM-YOLOv5s algorithm proposed in this paper are trained on the home-made PSO 2023 dataset, respectively, and the loss convergence curves and mAP curves during the training period are shown in Figure 8. The loss convergence curves of the two algorithms are compared in Figure 8a, and the loss convergence curves of the two algorithms are compared at mAP_0.5 in Figure 8b.



(**a**) (**b**)

Figure 8. Comparative analysis of model evaluation indicators. (**a**) Overall loss convergence curves; (**b**) mAP@0.5 Convergence curves.

From Figure 8a, it can be observed that both YOLOv5s and MFMAM-YOLOv5s have the highest convergence rate within 25 epochs. However, the algorithm model proposed in this paper demonstrates a faster rate of descent and convergence of the loss curve compared to the original YOLOv5s network model. Additionally, as shown in Figure 8b, both algorithm models demonstrate a similar upward trend in the first 20 epochs and stabilize around 290 epochs without any further increase. However, after 20 epochs, the mean average precision curve of the MFMAM-YOLOv5s algorithm model shows a significantly faster increase compared to the YOLOv5s algorithm model. In other words, the MFMAM-YOLOv5s algorithm model has a higher overall mAP value compared to the YOLOv5s algorithm model.

4.2. Comparison of Different Attention Mechanisms

To mitigate the interference caused by complex backgrounds and enhance the ability to focus on important features, this paper presents a hybrid attention mechanism that integrates channel attention and spatial attention. The spatial attention mechanism primarily focuses on the spatial dimension in the input data. It suppresses complex background features and enhances important features by assigning different weights. The channel attention mechanism primarily focuses on the channel dimension in the data. It learns which channels are more important for the task, emphasizes important feature information, and reduces the impact of background noise and redundant features.

To further validate the effect of the channel attention mechanism and spatial attention mechanism on the model, in this paper, experiments are conducted on the self-made PSO 2023 dataset to test the current mainstream Global Attention Module (GAM), Efficient Channel Attention (ECA), Multidimensional Collaborative Attention (MCA), and CBAM attention mechanisms. The experimental results are shown in Table 5.

Table 5. Comparison of the experimental results of different attention mechanisms.

Method	Volume/MB	Parameters/M	FLOPs/G	FPS/f·s^{-1}	mAP_0.5/%
YOLOv5s	14.4	7.04	16.0	303.0	94.3
GAM	17.1	8.77	17.2	416.7	95.1
ECA	13.9	7.08	16.2	434.8	95.0
MCA	16.7	7.82	16.8	439.4	95.6
CBAM	14.8	7.55	16.6	454.5	96.2

From the experimental results in Table 5, it can be observed that, compared to YOLOv5s without any attention mechanisms, the use of the GAM attention mechanism increased the mAP value by 0.8%. The model parameters and floating-point operations increased by 1.73 M and 1.2 G, respectively, while the model detection rate increased by 113.7 f·s^{-1}. Similarly, with the introduction of the ECA and MCA mechanisms, the mAP values increased by 0.7% and 1.3%, respectively. The model parameters and floating-point operations also increased, and the model detection rates were improved by 131.8 f·s^{-1} and 136.4 f·s^{-1}, respectively. However, upon introducing the CBAM attention mechanism, the model parameters and floating-point operations only increased by 0.51 M and 0.6 G, while the model detection rate improved to 454.5 f·s^{-1}, and the mAP value also increased by 1.9%. Hence, the CBAM hybrid attention mechanism demonstrates better performance in detecting pole-shaped obstacles in complex environments.

4.3. Ablation Experiment

To verify the effectiveness of the proposed multiscale feature hybrid attention algorithm, this experiment was conducted on the homemade PSO 2023 dataset for ablation experiments of the MFMAM-YOLOv5s algorithm. The experimental results are shown in Table 6.

Table 6. Results of ablation experiments.

MKS	CBAM	GFPN	Rep-GFPN	mAP_0.5/%	Volume/MB	FPS/f·s^{-1}
×	×	×	×	94.3	14.4	303.0
√	×	×	×	95.6	13.7	454.5
√	√	×	×	96.2	14.7	454.5
√	×	√	×	95.8	14.8	416.7
√	×	×	√	96.1	15.0	434.8
√	√	√	×	96.9	18.6	384.6
√	√	×	√	97.4	19.0	400.0

GFPN: original feature pyramid; Rep-GFPN: improved feature pyramid. ×: the module was not introduced; √: introduction of the module.

From the above experimental results, it is evident that utilizing the enhanced loss function results in enhancements in both the mAP value and the image detection rate when compared to the original model. Meanwhile, the size of the model has also been reduced. The main reason for this is that the improved loss function incorporates an optimal matching method. This method addresses the issue of the overlapping and mismatching of multiple bounding boxes in the original model. As a result, it significantly accelerates the matching process of predicted boxes and reduces the model size, making it more lightweight.

However, after solving the problem of optimal matching boxes, this paper introduces a hybrid attention mechanism to further improve the detection accuracy. This attention mechanism not only suppresses interference caused by complex environments and enhances the extraction of important features by allocating different weights spatially but also learns which channels are more important for the task. It focuses on important feature information, reduces background noise, and eliminates redundant features in the channel dimension. Through ablation experiments, it has further been proven that the incorporation of a hybrid attention mechanism allows the model to concentrate on significant features in both channels and mitigate the interference from intricate backgrounds in both channels.

To address the issue of multi-scale features, this paper introduces a multi-scale feature pyramid module. However, the original feature pyramid GFPN, due to its complex convolutions, increases the computational complexity and slows down the training and inference speed. Therefore, this paper utilizes an enhanced multi-scale feature pyramid known as Rep-GFPN. To evaluate the performance of the original GFPN and the enhanced Rep-GFPN, we conducted ablation experiments. From Table 6, it can be observed that the improved feature pyramid outperforms the original GFPN in terms of detection accuracy and speed. In summary, when compared to the YOLOv5s model, the enhanced model demonstrates a 3.1% increase in the mAP value and a 97 $f \cdot s^{-1}$ improvement in the detection speed. Although the model size has increased, the algorithm's overall performance is the best, fully meeting the detection requirements of ADAS systems.

4.4. Comparative Experiments of Different Algorithms

In order to assess the impact of algorithm enhancements, this paper conducted experimental verification on the self-made dataset PSO 2023 using the object detection algorithms YOLOv4, YOLOv4-tiny, Faster-RCNN, YOLOv5, YOLOv7, and YOLOv8. The experimental results are shown in Table 7.

Table 7. Comparative Results of Experiments with Different Models.

Model	Backbone	mAP_0.5/%	FPS/f·s^{-1}	Volume/MB
YOLOv4	CSPDarknet53	91.6	36.9	244.5
YOLOv4-tiny	CSPDarknet53-Tiny	83.9	153.9	22.5
Faster-RCNN	Resnet50	82.3	23.0	108.0
Faster-RCNN	VGG16	88.5	17.6	521.0
YOLOv5s	CSPDarknet53	94.3	303.0	14.4
YOLOv5m	CSPDarknet53	94.6	263.2	40.3
YOLOv5l	CSPDarknet53	95.3	156.3	88.6
YOLOv5x	CSPDarknet53	95.7	94.3	165.2
YOLOv7	CSPDarknet53	95.0	104.2	71.5
YOLOv8s	Darknet53	97.5	109.9	22.5
MFMAM-YOLOv5s (our)	CSPDarknet53	97.4	400.0	19.0
MFMAM-YOLOv5m (our)	CSPDarknet53	97.8	277.8	61.3
MFMAM-YOLOv5l (our)	CSPDarknet53	98.2	147.1	97.1
MFMAM-YOLOv5x (our)	CSPDarknet53	98.8	88.5	179.7

It is evident from the table that the two-stage object detection algorithm, Faster-RCNN, did not achieve higher accuracy despite using the Resnet50 backbone network. Instead, it resulted in a slower detection frame rate. YOLOv4-tiny and YOLOv5, commonly used lightweight algorithms, may have certain advantages in model deployment. However, they are not sufficiently outstanding in terms of detection accuracy and real-time performance for pole-shaped obstacles in complex environments. As a result, they struggle to meet the requirements of assisted driving perception technology. Although the YOLOv7 and YOLOv8 algorithms have a relatively good detection accuracy, the YOLOv7 model is larger and is not the optimal choice for deploying a perception system. Although the YOLOv8s model has a high detection accuracy, the image detection rate is not high. For an object detection system, it is necessary to ensure both accurate and speedy detection. Therefore, the method proposed in this article improves the mean Average Precision (mAP) to varying degrees in MFMAM-YOLOv5s, MFMAM-YOLOv5m, MFMAM-YOLOv5l, and MFMAM-YOLOv5x. In particular, MFMAM-YOLOv5s has significantly improved frames per second (FPS) and mean average precision (mAP) values compared to YOLOv5s and YOLOv7. Compared to the aforementioned models, the MFMAM-YOLOv5s model has a better balance between speed and accuracy, making it more suitable for deployment in assisted driving systems.

To visually evaluate the performance of the improved algorithm in terms of detection accuracy and robustness, this article selected the test result images of the YOLOv5s, YOLOv7, YOLOv8s, and MFMAM-YOLOv5s algorithms on the test dataset. These images are shown in Figure 9.

(**a**) (**b**) (**c**) (**d**)

Figure 9. Comparison Results Between Different Algorithms. (**a**) YOLOv5s; (**b**) YOLOv7; (**c**) YOLOv8s; (**d**) MFMAM-YOLOv5s.

From the comparative effect chart shown in Figure 9, it can be observed that the YOLOv5s network model exhibits a relatively high false detection rate and missed detection rate in environments with obstructions, rainy and snowy weather, and low visibility, although it is capable of detecting pole-like obstacles. The YOLOv7 and YOLOv8s network models also experience higher false detection and missed detection rates in complex environments with partial feature obstructions, as well as in rainy and snowy weather. However, compared to the aforementioned object detection algorithms, the improved

MFMAM-YOLOv5s network model is still capable of accurately detecting pole-like obstacles in complex environments.

4.5. Binocular Camera Target Ranging

The detection of pole-shaped obstacles not only requires more accurate identification and classification but also requires a sufficient ranging accuracy within a certain range. Therefore, this paper first uses the improved network model to obtain the coordinate information of the target; then, the coordinate information of the target and the stereo matching in the binocular ranging algorithm are correlated to obtain a more accurate measurement distance through the calculation of multiple information.

To further validate the accuracy and robustness of this method, real road test experiments were conducted. First, a section of an urban road is arbitrarily selected to detect road cones using the method in a complex environment; then, the center of light of the vehicle-mounted camera is taken as the origin, and a laser rangefinder is used to verify and read the real distance. At the same time, the error of the laser rangefinder is ±1.5 mm, which is in line with the measurement accuracy. Finally, the above experiments were repeated and recorded by constantly changing the distance. At the same time, to better demonstrate the error between the actual distance and the test distance, this paper compares and analyzes the actual distance, the test distance, the absolute error, and the relative error of the 80 groups of experimental results and draws a comparative statistical graph, as shown in Figure 10.

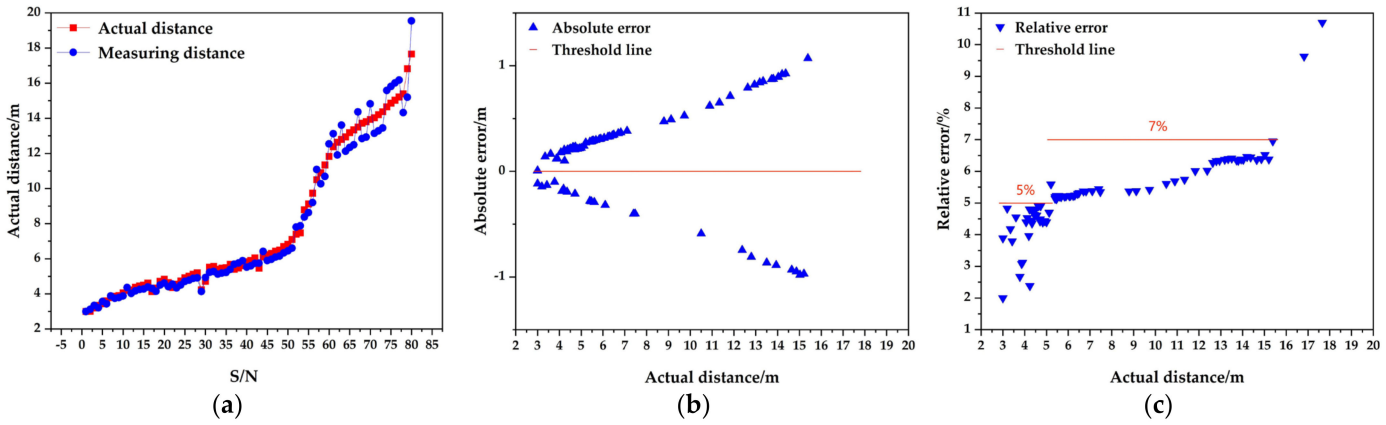


Figure 10. Statistical Comparison Chart of Ranging Experiment Results. (**a**) Comparison of actual distance and test distance; (**b**) Absolute error statistics chart; (**c**) Relative error statistics.

From Figure 10a, it can be observed that the output distance of the pole-shaped obstacle detection system fluctuates above and below the true distance. Meanwhile, as the distance increases, the fluctuation also increases. As for Figure 10b, it shows the absolute error, which clearly increases with distance. In other words, the absolute error value is proportional to the distance. From Figure 10c, it can be visually seen that the relative error within a range of 6 m is less than 5%. The relative error within a range of 3–15 m is less than 7%, which meets the accuracy requirements for distance measurement. In conclusion, as the distance increases, the measurement accuracy of the pole-shaped obstacles gradually decreases, but the overall detection accuracy fully meets the requirements.

The above experiments were all conducted on static straight roads. However, if the vehicle is traveling in a dynamic scene, it will have a certain impact on the de-detection accuracy and real-time performance. As the vehicle is moving rapidly, the feature information of the pole obstacle is also changing, such as the length, size, angle, etc. The limitation of binocular camera hardware and the vibration of the vehicle body will cause some difficulties in the feature extraction of rod-shaped obstacles. Therefore, to further validate the accuracy and robustness of the improved algorithm in dynamic situations, a

section of the curve was directly selected for testing in this experiment. If the algorithm can accurately detect pole-shaped obstacles while the vehicle is making a turn, it would serve as a good validation for road detection in dynamic scenarios. According to the regulations of the Road Traffic Safety Law [37], the maximum speed for vehicles to make a turn is 30 km/h. Therefore, this study conducted experiments using the maximum turning speed. In this paper, 10 turning tests were conducted with the maximum turning speed of 30 km/h throughout the experiment, and 20 sets of experimental data were obtained, as shown in Table 8.

Table 8. Results of the turn test experiment.

S/N	Speed/km·h^{-1}	Value/m	Output/m	Relative Error/%	FPS/f·s^{-1}
01	30	3.125	2.998	4.064	31.1
02	30	3.421	3.267	4.502	32.3
03	30	3.028	3.146	3.897	26.8
04	30	4.127	4.320	4.677	28.0
05	30	4.335	4.545	4.844	31.3
06	30	4.870	4.612	5.298	27.4
07	30	5.465	5.785	5.855	26.5
08	30	5.233	5.530	5.676	29.6
09	30	5.674	5.316	6.309	30.8
10	30	6.130	6.517	6.313	27.7
11	30	6.422	6.005	6.493	31.5
12	30	6.700	7.160	6.866	26.8
13	30	7.152	7.645	6.893	27.6
14	30	8.400	8.986	6.976	26.1
15	30	9.710	10.390	7.003	28.3
16	30	10.270	9.540	7.108	31.6
17	30	11.295	12.121	7.313	26.4
18	30	12.430	13.395	7.763	28.3
19	30	13.500	14.564	7.881	28.1
20	30	14.664	13.501	7.931	26.7

From the test experimental results in Table 8, it can be observed that the detection accuracy of the objects is somewhat affected due to the rapid changes in target features during the turning process. However, as seen from the above test results, although the detection accuracy of pole-shaped obstacles is slightly reduced during turns, the overall detection performance still meets the requirements of object detection. Therefore, the fast-turning experiment further validates the effectiveness of detection on dynamic straight roads.

5. Conclusions

In this paper, the binocular camera is first stereo-calibrated, and the obtained internal and external parameters are stereo-corrected for the real-time acquired images, respectively. Subsequently, the targets in the acquired images are recognized and classified using the improved network model. Then, the coordinate information of the target frame was correlated with the stereo matching to obtain more accurate information about the kind and distance between the camera and the target. Finally, the accuracy and robustness of the target detection algorithm were verified through experiments, respectively. The experimental results show that the mAP_0.5 value of the proposed method is 97.4%, which is 3.1% higher than that of the original algorithm YOLOv5s, the detection rate is 97 f·s^{-1} higher than that of the original algorithm, and the ranging error of the pole obstacle in a complex environment is also less than 7%. The method not only recognizes the target quickly for classification but also accurately measures the distance between the camera and the target.

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

Funding: This research was funded by the Beijing Smarter Eye Technology Co., Ltd. cooperation project, grant numbers 2200010047, 2100010024, and 1900010008.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the authorized access to data.

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

References

1. Bokaba, T.; Doorsamy, W.; Paul, B.S. Comparative study of machine learning classifiers for modelling road traffic accidents. *Appl. Sci.* **2022**, *12*, 828. [CrossRef]
2. Brijs, T.; Mauriello, F.; Montella, A.; Galante, F.; Brijs, K.; Ross, V. Studying the effects of an advanced driver-assistance system to improve safety of cyclists overtaking. *Accid. Anal. Prev.* **2022**, *174*, 106763. [CrossRef] [PubMed]
3. Bosurgi, G.; Pellegrino, O.; Ruggeri, A.; Sollazzo, G. The Role of ADAS While Driving in Complex Road Contexts: Support or Overload for Drivers. *Sustainability* **2023**, *15*, 1334. [CrossRef]
4. Wang, Y.; Liu, H.; Chen, N.J.A.S. Vehicle detection for unmanned systems based on multimodal feature fusion. *Appl. Sci.* **2022**, *12*, 6198. [CrossRef]
5. Badrloo, S.; Varshosaz, M.; Pirasteh, S.; Li, J. Image-based obstacle detection methods for the safe navigation of unmanned vehicles: A review. *Remote Sens.* **2022**, *14*, 3824. [CrossRef]
6. Huang, Z. Semantic road segmentation based on adapted Poly-YOLO. In Proceedings of the 3rd International Conference on Signal Processing and Machine Learning (CONF-SPML), Oxford, UK, 18 August 2023; pp. 012–015. [CrossRef]
7. Mijwil, M.M.; Aggarwal, K.; Doshi, R.; Hiran, K.K.; Gök, M. The Distinction between R-CNN and Fast RCNN in Image Analysis: A Performance Comparison. *Asian J. Appl. Sci.* **2022**, *10*, 429–437. [CrossRef]
8. Arora, N.; Kumar, Y.; Karkra, R.; Kumar, M. Automatic vehicle detection system in different environment conditions using fast R-CNN. *Multimed Tools Appl.* **2022**, *81*, 18715–18735. [CrossRef]
9. Liu, T.; Stathaki, T. Faster R-CNN for robust pedestrian detection using semantic segmentation network. *Front. Neurorobotics* **2018**, *12*, 64. [CrossRef]
10. Lai, K.; Zhao, J.; Liu, D.; Huang, X.; Wang, L. Research on pedestrian detection using optimized mask R-CNN algorithm in low-light road environment. In Proceedings of the 9th Global Conference on Materials Science and Engineering (CMSE), Kyiv, Ukraine, 15–17 January 2021; pp. 012–057. [CrossRef]
11. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
12. Dhall, A.; Dai, D.; Van Gool, L. Real-time 3D traffic cone detection for autonomous driving. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 494–501. [CrossRef]
13. He, Y.; Liu, Z. A feature fusion method to improve the driving obstacle detection under foggy weather. *IEEE Trans. Transp. Electrif.* **2021**, *7*, 2505–2515. [CrossRef]
14. Liu, Z.; Zhao, S.; Wang, X. Research on driving obstacle detection technology in foggy weather based on GCANet and feature fusion training. *Sensors* **2023**, *23*, 2822. [CrossRef] [PubMed]
15. Pan, X.; Yi, Z.; Tao, J. The research on social distance detection on the complex environment of multi-pedestrians. In Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; pp. 763–768. [CrossRef]
16. 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]
17. Shuai, G.; Wenlun, M.; Jingjing, F.; Zhipeng, L. Target recognition and range-measuring method based on binocular stereo vision. In Proceedings of the 2020 4th CAA International Conference on Vehicular Control and Intelligence (CVCI), Hangzhou, China, 18–20 December 2020; pp. 623–626. [CrossRef]
18. Du, S.; Zhang, B.; Zhang, P.; Xiang, P. An improved bounding box regression loss function based on CIOU loss for multi-scale object detection. In Proceedings of the 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, 16–18 July 2021; pp. 92–98. [CrossRef]
19. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
20. Chen, Y.; Gangbo, W.; Georgiou, T.T.; Tannenbaum, A. On the matrix Monge–Kantorovich problem. *Eur. J. Appl. Math.* **2020**, *31*, 574–600. [CrossRef]

21. Lehmann, T.; Von Renesse, M.-K.; Sambale, A.; Uschmajew, A. A note on overrelaxation in the Sinkhorn algorithm. *Optim. Lett.* **2021**, *16*, 2209–2220. [CrossRef]
22. Juraev, G.; Rakhimberdiev, K. Mathematical modeling of credit scoring system based on the Monge-Kantorovich problem. In Proceedings of the 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 1–4 June 2022; pp. 1–7. [CrossRef]
23. Bogachev, V. Kantorovich problem of optimal transportation of measures: New directions of research. *Russ. Math. Surv.* **2022**, *77*, 769–817. [CrossRef]
24. Wang, X.; Pan, Z.; Gao, H.; He, N.; Gao, T. An efficient model for real-time wildfire detection in complex scenarios based on multi-head attention mechanism. *J. Real-Time Image Proc.* **2023**, *20*, 66. [CrossRef]
25. Lee, D.; Jang, K.; Cho, S.Y.; Lee, S.; Son, K. A Study on the Super Resolution Combining Spatial Attention and Channel Attention. *Appl. Sci.* **2023**, *13*, 3408. [CrossRef]
26. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. Damo-yolo: A report on real-time object detection design. *arXiv* **2022**, arXiv:2211.15444.
27. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Cham, Switzerland, 6 October 2018; pp. 3–19. [CrossRef]
28. Guo, Y.; Zeng, Y.; Gao, F.; Qiu, Y.; Zhou, X.; Zhong, L.; Zhan, C. Improved YOLOv4-CSP algorithm for detection of bamboo surface sliver defects with extreme aspect ratio. *IEEE Access* **2022**, *10*, 29810–29820. [CrossRef]
29. Lu, P.; Liu, Q.; Guo, J. Camera calibration implementation based on Zhang Zhengyou plane method. In *Proceedings of the 2015 Chinese Intelligent Systems Conference*; Springer: Berlin/Heidelberg, Germany; pp. 29–40. [CrossRef]
30. Zhang, P.; Liu, Z. Research on Binocular Stereo Vision Ranging Based on Improved YOLOv5s. In Proceedings of the 2023 5th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP), Chengdu, China, 19–21 May 2023; pp. 1242–1246. [CrossRef]
31. Tang, H.; Sun, W. A discussion of the Bouguer correction. *Pure Appl. Geophys.* **2021**, *178*, 3543–3557. [CrossRef]
32. Huang, H. Research on binocular vision ranging based on YOLO algorithm and stereo matching algorithm. In Proceedings of the Second International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2023), Xishuangbanna, China, 2 May 2023; pp. 274–279. [CrossRef]
33. Deng, C.; Liu, D.; Zhang, H.; Li, J.; Shi, B. Semi-Global Stereo Matching Algorithm Based on Multi-Scale Information Fusion. *Appl. Sci.* **2023**, *13*, 1027. [CrossRef]
34. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
35. Rebuffi, S.-A.; Gowal, S.; Calian, D.A.; Stimberg, F.; Wiles, O.; Mann, T.A. Data augmentation can improve robustness. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29935–29948. [CrossRef]
36. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujscie, Poland, 9–12 May 2018; pp. 117–122. [CrossRef]
37. State Council of the People's Republic of China. *Regulation on the Implementation of the Road Traffic Safety Law of the People's Republic of China. Chapter IV, Road Access Regulations*; People's Public Security University of China Press: Beijing, China, 2004.