An Enhanced Deep Learning Model for Obstacle and Traffic Light Detection Based on YOLOv5

Zhenwei Li *, Wei Zhang and Xiaoli Yang

School of Medical Technology and Engineering, Henan University of Science and Technology, Luoyang 471023, China; 200320221493@stu.haust.edu.cn (W.Z.); yangxiaoli39@163.com (X.Y.)
* Correspondence: 9903137@haust.edu.cn

Abstract: Timely detection of dynamic and static obstacles and accurate identification of signal lights using image processing techniques is one of the key technologies for guidance robots and is a necessity to assist blind people with safe travel. Due to the complexity of real-time road conditions, current obstacle and traffic light detection methods generally suffer from missed detection and false detection. In this paper, an improved deep learning model based on YOLOv5 is proposed to address the above problems and to achieve more accurate and faster recognition of different obstacles and traffic lights that the blind may encounter. In this model, a coordinate attention layer is added to the backbone network of YOLOv5 to improve its ability to extract effective features. Then, the feature pyramid network in YOLOv5 is replaced with a weighted bidirectional feature pyramid structure to fuse the extracted feature maps of different sizes and obtain more feature information. Finally, a SIoU loss function is introduced to increase the angle calculation of the frames. The proposed model’s detection performance for pedestrians, vehicles, and traffic lights under different conditions is tested and evaluated using the BDD100K dataset. The results show that the improved model can achieve higher mean average precision and better detection ability, especially for small targets.

Keywords: obstacle and traffic light detection; deep learning model; YOLOv5; coordinate attention; bidirectional feature pyramid network

1. Introduction

At present, there are more than 2.2 billion visually impaired people worldwide, which is a very large vulnerable group. The vast majority of visually impaired people carry out their daily lives within a limited area because they encounter many “obstacles” when traveling, such as pedestrians, stalled vehicles, road signs, and safety facilities. These obstacles can injure them and may, in severe cases, lead to their deaths. In recent years, researchers have continuously attempted to use guided robots to help blind people navigate and improve their quality of life. Multi-form and multifunctional guide robots continue to enter the public eye, among which the most representative one is mobile guidance robots [1–3]. Timely detection of dynamic and static obstacles and accurate identification of signal lights is one of the key technologies for guidance robots and is a necessity to assist blind people with safe travel [4]. There are many obstacle detection techniques in the area of robot guiding, for example, ultrasonic ranging, infrared ranging, impact sensing [1], and image sensing [5]. Obstacle detection based on ultrasonic ranging and infrared ranging can easily detect the location of an obstacle, but these methods are easily affected by rain and fog and also cannot obtain the actual image of the detection target. Impact sensors can only detect objects that are very close. These shortcomings limit the development and practical application of these three types of obstacle detection techniques.

With the development of neural-network-related technologies, obstacle detection and target identification based on image sensing using deep learning has made a breakthrough. Obstacle detection algorithms based on deep learning include two categories: the two-stage...
algorithm of the R-CNN series and the one-stage algorithm of the YOLO series [6]. The R-CNN algorithms can be divided into two stages. The first stage mainly extracts the target area, and in the second stage, target detection is performed. Two-stage obstacle detection algorithms include the SPP-Net model, Fast R-CNN, and, further, Faster R-CNN [7–9]. Since the obstacle detection algorithms of the R-CNN family need to traverse all candidate frames when detecting obstacles, which takes a long time, they are unable to meet the needs of real-time applications. The YOLO algorithm proposed by Redmon is a one-stage obstacle detection algorithm that directly extracts the feature information of the image through the neural network model and outputs the result at the last layer of the model [10]. The detection efficiency of the YOLO algorithm is higher than that of the R-CNN series of target detection algorithms, but in its detection process, some target features may be lost, which results in low accuracy for this algorithm. The one-stage-based obstacle detection models include YOLOv2, YOLOv3, YOLOv4, YOLOv5, YOLOv6, YOLOv7, and SSD network structures [11–16]. In terms of obstacle detection for guiding the blind, Qiu Xuyang [17] adopted the Denset network to improve YOLOv3 and combined it with a stereo matching model to realize obstacle detection in traffic scenes. Duan Zhongxing [18] used the deep learning model YOLOv4 in blind path obstacle recognition, introduced asymmetric convolution and other modules to improve the network, and achieved higher accuracy in a practical scenario test. Wang Weifeng [19] adopted the method of increasing the receptive field to realize target recognition on the SSD network and improved the detection speed. Jiang and Yu proposed a pedestrian detection method based on XGBoost, which optimizes the XGBoost model with a genetic algorithm [20]. Xin Huang realized the recognition of crowded pedestrians using a novel Paired-Box Model (PBM) [21]. Ouyang realized real-time traffic light detection on the road through a CNN [22]. Vasiliki Balaska designed a system for generating enhanced semantic maps to detect obstacles in ground and satellite images [23]. Avinash Padmakar Rangari used YOLOv7 to design an intelligent traffic light system that realized rapid traffic by detecting vehicles, pedestrians, and obstacles in the traffic road [24]. Yan Zhou proposed an improved Faster-RCNN obstacle detection algorithm to recognize small targets and occluded targets in automatic driving scenarios [25]. Mazin Hnewa has designed a new multi-scale domain adaptive MS-DAYOLO for object detection [26]. Shuiye Wu proposed a YOLOX based network model for multi-scale object detection tasks in complex scenes [27]. Vicent ortiz castelló In order to help reduce accidents in the advanced driver assistance system, the original Leaky ReLU convolution activation function in the original YOLO implementation is replaced by the cutting-edge activation function in the YOLOv4 network to improve the detection performance [28]. When identifying obstacles in front of the blind, the obstacle detection algorithms mentioned above often have missed detection and false detection due to the diverse types of obstacles and complex conditions, such as occlusion, low contrast between target and background, and small target size.

In this paper, an improved obstacle detection model based on YOLOv5 is proposed to address the above problems and to achieve more accurate and faster recognition of different obstacles and traffic lights that the blind may encounter. The main contributions of this paper are as follows:

1. An improved deep learning model based on YOLOv5 is proposed. A coordinate attention layer is added to the backbone network of YOLOv5 to improve its ability to extract effective features in the target image. The feature pyramid network in YOLOv5 is replaced with a weighted bidirectional feature pyramid structure to fuse the extracted feature maps of different sizes and obtain more feature information. A SIoU loss function is introduced to increase the angle calculation of the frame, improve the detection speed of the frame regression, and improve the mean average precision.

2. The proposed model’s detection performance for pedestrians, vehicles, and traffic lights under different conditions is tested and evaluated using the BDD100K database [29]. The results show that the improved model can achieve higher mean average precision and better detection ability, especially for small targets.
2. Methodology

In YOLOv5, the CSPDarknet53 backbone network is used to extract the image features and includes Conv-Batchsize-SiLU (CBS), Spatial Pyramid Pooling—Fast (SPPF), and cross-stage partial network modules. The CBS structure is composed of convolution, batch normalization, and the SiLu activation function. The main role of SPPF is to extract high-level features and then fuse them. During the fusion process, the maximum-set operation is used several times to extract as many high-level semantic features as possible. Cross-Stage Partial networks (CSPs) used in backbone networks are added to the residual structure to more effectively extract high-level features [30]. In addition, CSP modules without residual structures are also used in feature fusion modules. The cross-stage network can describe the change in the image gradient in the feature graph, which reduces the network parameters and maintains the speed without losing precision [31]. The prediction part of the YOLOv5 model uses feature maps of three scales to generate prediction boxes for targets in the image and uses a non-maximum suppression method to obtain the box that is most similar to the real box [32].

In this paper, we propose the improved YOLOv5 algorithm to solve problems such as missed detection and false detection of different obstacles and traffic lights under different conditions. By improving the backbone network and replacing the feature fusion network and loss function, the network model is more suitable for road-object detection and recognition in practical application scenarios. The structure of the improved YOLOv5 model is shown in Figure 1.

Figure 1. The improved YOLOv5 network structure.

2.1. Coordinate Attention Module

In recent years, the effectiveness of the attention mechanism in computer vision has been proven and has been widely used in target classification [33], detection [34], and segmentation [35]. The attention mechanism can help the network model pay more attention to the feature and location information of the region of interest as well as improve the performance of the model. It is worth noting that the use of attention mechanisms requires a lot of computing to implement, which means more time with larger computing
devices. The coordinate attention (CA) mechanism can solve the computational overhead caused by most of the attention mechanisms, and it can embed the location information into the channel attention to help the model extract the features containing the channel information, direction information, and location information [36,37].

The feature map carries out global average pooling, convolution, and nonlinear activation functions along the dimensions of width and height, adding the location information of features to the channel attention. The feature plus coordinate attention calculation is shown in Equation (1).

\[ y_c(i, j) = x_c(i, j) \ast g_h^c(i, j) \ast g_w^c(i, j), \]

where \( g_h^c \) and \( g_w^c \) are the weights obtained in the width and height directions of the feature map, respectively, \( x_c \) is the input feature, and \( y_c \) is the output feature.

The backbone network of the YOLOv5 algorithm uses the CSP-Darknet53 network structure for feature extraction, which is based on the introduction of the CSPNet network structure on top of Darknet53. To further improve the feature extraction capability of the backbone network and ensure a reduction in computation and the improvement of the inference speed without a decrease in detection accuracy, a coordinate attention network structure is added. The CA module is shown in the upper right of Figure 1.

### 2.2. Bidirectional Feature Pyramid Network model

The original YOLOv5 uses the classical approach of combining a feature pyramid network (FPN) and path aggregation network (PANet) for feature fusion [38,39]. The FPN algorithm extracts deep semantic features from the top down, while the PANet algorithm obtains target location information from the bottom up. By integrating the information obtained from these two algorithms, the information about features is increased and the sensitivity of the network is improved. Compared with the classical approach of combining FPN and PANet, the bidirectional feature pyramid network (BiFPN) adds cross-layer connections and removes nodes with only one input edge [40]. Multiple operations can be performed on the same layer to realize feature fusion at a higher level. BiFPN is simpler and faster in multi-scale feature fusion than PANet because it requires fewer parameters and calculations. The BiFPN structure can make the prediction network more sensitive to objects with different resolutions and improve the performance of the detection network. Therefore, this paper adopts the BiFPN structure to ensure that the model can enhance the semantic and localization information of the features and improve the detection rate at the same time. The BiFPN structure is shown in Figure 2.
2.3. Loss Function

The original YOLO5 model uses CIoU as its loss function [41]. In boundary box regression calculations, the CIoU loss function adds the distance between the center point of the real frame and the predicted frame, the vertical-to-horizontal ratio, and the area of the overlapping part of the two frames, but it does not consider the regression direction between the real frame and the predicted frame, resulting in slow convergence speed and low efficiency. The SIoU loss function increases the vector angle between required regressions, redefines the penalty index, and adds the angle to the calculation of the distance between the real frame and the prediction frame [42]. Compared to CIoU, the SIoU loss function can achieve faster convergence speed and higher computational efficiency. Therefore, this paper introduces the SIoU to replace the CIoU.

The SIoU loss function is as follows:

\[ L_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2}, \]  

where \( \Delta \) and \( \Omega \) represent the distance cost and the shape cost, respectively. \( \Delta \) is defined as:

\[ \Delta = \sum_{t=x,y} (1 - e^{-\gamma \rho_t}), \]  

where

\[ \rho_t = \begin{cases} \frac{(\hat{b}_{tx} - b_{cx})^2}{\hat{c}_w}, & t = x \\ \frac{(\hat{b}_{ty} - b_{cy})^2}{\hat{c}_h}, & t = y \end{cases} \]  

\[ \gamma = 2 - \Lambda, \]  

where \( \Lambda \) represents the angle cost:

\[ \Lambda = 1 - 2 \sin^2(\arcsin(x)) - \frac{\pi}{4}; \]  

\( \Omega \) is defined as:

\[ \Omega = \sum_{t=w,h} (1 - e^{-\omega_t})^\theta. \]

The SIoU loss function greatly improves the training and reasoning of the target detection algorithm, realizes faster convergence, and has better performance in reasoning, which can help the model converge faster and more accurately.

3. Experiment Results and Discussion

3.1. BDD100K Dataset

The dataset used in this experiment is derived from the road detection dataset BDD100K released by the University of California (Berkeley) in 2018 and contains a collection of 100,000 images of the streets of several different cities. There are common static and dynamic targets, such as traffic lights, pedestrians, motorcycles, and cars in the images. The images in the dataset have different levels of clarity to ensure their diversity. Considering the main obstacles encountered by blind people while traveling are pedestrians, vehicles, and traffic lights, we select these three different obstacles as the detection targets in the experiment and generate the training set and test set at random in a ratio of 8:2.
3.2. Evaluation Metrics

In this paper, IOU, Precision, Recall, and mAP50 are used to assess the performance. IOU refers to the area overlap ratio of the obtained detection and the ground truth boxes and can be expressed as:

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

(8)

Precision and Recall are defined as:

$$ \text{Precision} = \frac{TP}{TP + FP} $$

(9)

$$ \text{Recall} = \frac{TP}{TP + FN} $$

(10)

where $TP$, $TN$, and $FN$ stand for true positives, true negatives, and false negatives, respectively. The Mean Average Precision ($mAP$) can be calculated as follows:

$$ mAP = \frac{1}{N} \sum_{i=0}^{N} \left( \text{Recall} = \frac{i}{N} \right) $$

(11)

where $N$ denotes the number of classes. AP50 represents the Mean Average Precision at $IoU = 0.5$ and is one of the main indicators used to compare the performance of different models.

3.3. Results and Discussion

All experiments in this paper were conducted in the same configuration environment and training environment configuration: AMD 5950X CPU, 64GB RAM, and NVIDIA GeForce RTX3090 24G GPU. The PyTorch1.12 framework and Anaconda Python3.7 interpreter built the network framework. The iteration period was set to 300 times for training, and the accuracy and average accuracy of the model were calculated.

3.3.1. Detection Results

We evaluated the results of the proposed method and six other representative methods using the mAP and AP50 indicators on the same BDD100K dataset. Table 1 shows the Precision and AP50 of the detection results of the three different detection categories, which are pedestrians, vehicles, and traffic lights, with the proposed model. It can be seen from Table 1 that vehicles have the highest detection accuracy at 84.60%, while the detection accuracy of traffic lights is the lowest at only 69.20%. The AP50s of the three categories are 68.10%, 80.60%, and 55.20%, respectively. Similarly, vehicles have the highest AP50, while traffic lights have the lowest. After analyzing the dataset, we found that the number of yellow traffic lights was only 510, which is 1/200 of the number of vehicle labels, which affects the overall detection accuracy of the traffic lights. Meanwhile, the traffic lights in the images are much smaller than the pedestrians and vehicles, which, to some extent, poses difficulties for their detection. In brief, insufficient training data and smaller sizes in the images are two main factors that affect the accuracy of traffic light detection.

Table 1. The Precision and AP50 of the three different detection categories.

<table>
<thead>
<tr>
<th>Detection Category</th>
<th>Precision (%)</th>
<th>AP50 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pedestrians</td>
<td>75.10</td>
<td>68.10</td>
</tr>
<tr>
<td>vehicles</td>
<td>84.60</td>
<td>80.60</td>
</tr>
<tr>
<td>traffic lights</td>
<td>69.30</td>
<td>55.20</td>
</tr>
</tbody>
</table>

The computational complexity of the CA module was evaluated through comparative experiments. In the experiments, a Convolutional Block Attention Module (CBAM),
Squeeze Excitation (SE), and a CA module were added to the YOLOv5 model in turn, and then their parameters, GFLOPs, and mAP were calculated. Table 2 gives the comparison results. It can be seen from the table that the CA module in this paper results in higher predictive performance with less computational cost compared with CBAM and SE.

Table 2. Comparison of computational complexity between different attention mechanisms. (N = 2, Epoch = 50).

<table>
<thead>
<tr>
<th>Attention Mechanism</th>
<th>Parameters</th>
<th>GFLOPs</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>7.12 × 10^6</td>
<td>16.0</td>
<td>62.95</td>
</tr>
<tr>
<td>CBAM</td>
<td>7.13 × 10^6</td>
<td>16.1</td>
<td>62.70</td>
</tr>
<tr>
<td>CA</td>
<td>7.10 × 10^6</td>
<td>16.0</td>
<td>63.05</td>
</tr>
</tbody>
</table>

Table 3 gives the AP50 and mAP of the proposed method and six other classical objection detection algorithms. For AD-Faster-RCNN [25], SSD [16], YOLOv4-416 [28], and YOLOv4, as there are no detection data for traffic lights, only their pedestrian and vehicle detection results are provided, and the number of their detection categories N = 2. For MS-DAYOLO [26], YOLOv5, YOLOv6, YOLOv7-tiny [27], and the method in this paper, N = 3. From Table 3, the AP50s of the proposed model for pedestrian, vehicle, and traffic light detection reach 68.1%, 80.6%, and 55.2%, respectively, which are the best performances in the comparative experiments. Compared with the original YOLOv5 model, the AP50s of pedestrian, vehicle, and traffic light detection of our method has been increased by 11.9%, 4.6%, and 8.17%, respectively, and the mAP increases by 8.23%, demonstrating the effectiveness of the improved model. For YOLOv5, YOLOv6, YOLOv7-tiny, and the method in this paper, our model has the highest computational reasoning time, but it is still real-time, reaching 4.7 ms, which is an increase of 1.3 ms compared with the YOLOv5 model. It is wise to trade time for accuracy without compromising the real-time requirements of the model. Higher-precision detection models can be more capable of sensing the environment.

Table 3. Comparison of detection results of nine different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP (%)</th>
<th>AP50 (%)</th>
<th>Inference Time/ms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>People</td>
<td>Car</td>
</tr>
<tr>
<td>AD-Faster-RCNN</td>
<td>56.15</td>
<td>62.00</td>
<td>50.30</td>
</tr>
<tr>
<td>SSD</td>
<td>45.39</td>
<td>44.33</td>
<td>46.46</td>
</tr>
<tr>
<td>MS-DAYOLO</td>
<td>55.70</td>
<td>45.37</td>
<td>73.74</td>
</tr>
<tr>
<td>YOLOv4-416</td>
<td>60.86</td>
<td>50.32</td>
<td>71.40</td>
</tr>
<tr>
<td>YOLOv4</td>
<td>60.45</td>
<td>51.70</td>
<td>69.20</td>
</tr>
<tr>
<td>YOLOv5</td>
<td>66.10</td>
<td>56.20</td>
<td>76.00</td>
</tr>
<tr>
<td>YOLOv6</td>
<td>53.00</td>
<td>47.00</td>
<td>71.20</td>
</tr>
<tr>
<td>YOLOv7-tiny</td>
<td>56.03</td>
<td>51.70</td>
<td>72.60</td>
</tr>
<tr>
<td>Ours</td>
<td>67.97</td>
<td>68.10</td>
<td>80.60</td>
</tr>
</tbody>
</table>

3.3.2. Detection Performance under Special Circumstances

To test the performance of the model under special circumstances, such as occlusion, low contrast between target and background, and small target size, three different images for each type of detection target are randomly selected, and each image to be detected represents a situation. The detection results are shown in Figures 3–5, and some details are enlarged using yellow rectangular boxes to better carry out subjective visual evaluation. The first column (Figures 3a,d,h–5a,d,h) in each set of images represents the image to be detected, the second column (Figures 3b,e,i–5b,e,i) represents the detection results of the original YOLOv5 model, and the third column (Figures 3c,f,j–Figure 5c,f,j) represents the detection results of the improved model.
Figure 3 shows the comparison of vehicle detection results between the original YOLOv5 model and our model. The first row of images shows the detection ability of the proposed model for occluded targets. From Figure 3b, it can be seen that the original YOLOv5 mistakenly identified each part of the two vehicles in the marked box on the left side of Figure 3b as one vehicle, while in Figure 3c, this error detection has been corrected. The second row of images shows the detection performance under low contrast between the background and the target. In Figure 3e, the gray and black car in the yellow box was not successfully detected, while in Figure 3f, it was accurately identified. The third row of image illustrates the comparison of detection capabilities for small targets: in this case, vehicles in the distance. Figure 3i is the detection result of the original YOLOv5; the vehicle in the distance of the identification box was not recognized, but it was correctly detected in Figure 3j using our model. Figure 4 shows the pedestrian detection results using the original YOLOv5 model and our model. The first row compares the detection results under normal conditions (good image quality), the second row provides the detection comparison of pedestrians in shadows (low image contrast), and the third row illustrates the detection effect of distant pedestrians (small targets). In Figure 4b, one pedestrian was detected as two pedestrians; in Figure 4e, the vehicle and background in the yellow box were incorrectly detected as pedestrians; in Figure 4i, only one of the two pedestrians in the yellow box was detected. All the false and missed detections in Figure 4b,e,i using the original YOLOv5 model were corrected with our model, which can be seen from Figure 4c,f,j.
Figure 4. Comparison of pedestrian detection results between the original YOLOv5 model and our model. (a,d,h) are the images to be detected; (b,e,i) are the detection effect of YOLOv5; (c,f,j) are the detection effect of the improved model.

Accurately detecting traffic lights at a distance can help guidance robots plan their travel routes reasonably, thereby saving travel time for blind people. The main problem in long-distance traffic light detection is the missed and false detection caused by the small size of traffic lights in the image. Therefore, this paper takes the detection of traffic lights as one of the factors to evaluate the performance of the proposed model. Figure 5 compares the long-distance traffic light detection performance between the original YOLOv5 model and our model. In Figure 5a, there are two small green lights in the yellow box, only one of which was recognized with the YOLO5 model (as shown in Figure 5b). In Figure 5c, both green lights were detected using the proposed model. In the second row, the original YOLO5 model mistakenly detected the yellow signal light in Figure 5d as red (Figure 5e), while the model this paper proposes did the right thing, as shown in Figure 5f. In the third row, the improved model successfully identified the red traffic light at the next intersection that appears in the identification box (Figure 5j), while the YOLO5 model did not (Figure 5i).

From the above analysis, it can be seen that the improved YOLOv5 model has the performance of detecting targets more accurately in complex environments and can effectively detect targets.

3.3.3. Ablation Experiments

Ablation experiments were carried out to assess the contributions of the three modules to the performance of the improved model. The CA, BIFPN, and SIoU were removed from the proposed model separately. The experiments are designed as follows:

- Experiment 1: Remove SIoU, keep CA and BiFPN;
- Experiment 2: Remove BiFPN, keep CA and SIoU;
- Experiment 3: Remove CA, keep BiFPN and SIoU;
- Experiment 4: Keep all three modules.
The mAPs and AP50s of the four experiments are given in Table 4. It can be easily seen from the table that each innovative module in this paper contributes to the whole performance and decreases the mean average precision when removed. The CA module has the greatest impact on the proposed model. This is because coordinate attention can improve the model’s ability to filter important features, increase the proportion of effective features, and improve the model’s feature expression ability, thereby improving the model’s detection performance. The ablation experiment results show that the combination of the CA, weighted BiFPN, and SIoU loss function can effectively improve the performance of the model for object recognition and improve the model’s performance as a whole.

![Figure 5](image)

**Figure 5.** Comparison of long-distance traffic light detection results between the original YOLOv5 model and our model. (a,d,h) are the images to be detected; (b,e,i) are the detection effect of YOLOv5; (c,f,j) are the detection effect of the improved model.

**Table 4.** Ablation experiment results.

<table>
<thead>
<tr>
<th>Model</th>
<th>CA</th>
<th>BiFPN</th>
<th>SIoU</th>
<th>mAP (%)</th>
<th>AP50 (%)</th>
<th>Pedestrians</th>
<th>Vehicles</th>
<th>Traffic Lights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>√</td>
<td>√</td>
<td>×</td>
<td>65.50</td>
<td>64.60</td>
<td>80.30</td>
<td>51.60</td>
<td></td>
</tr>
<tr>
<td>Experiment 2</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>61.38</td>
<td>58.20</td>
<td>77.40</td>
<td>48.53</td>
<td></td>
</tr>
<tr>
<td>Experiment 3</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>60.97</td>
<td>53.70</td>
<td>76.90</td>
<td>48.20</td>
<td></td>
</tr>
<tr>
<td>Experiment 4</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>67.96</td>
<td>68.10</td>
<td>80.60</td>
<td>55.17</td>
<td></td>
</tr>
</tbody>
</table>

4. Conclusions

This paper proposed an improved deep learning model based on YOLOv5 for static and dynamic obstacles and traffic light detection. Considering the characteristics of obstacles (targets) that blind people may encounter in real-time when traveling, three different
modules—CA, BiFPN, and SLoU loss function—were introduced into the model to improve its detection ability. Detection precision tests, tests of detection performance under special conditions, and ablation experiments were conducted. The results of detection precision tests showed that the AP50s of pedestrian, vehicle, and traffic light detection of our method increased by 11.9%, 4.6%, and 8.17%, respectively, compared with the original YOLOv5 model, and the mAP increased by 8.23%, demonstrating the effectiveness of the improved model. Tests of detection performance under special conditions showed that the proposed model has the performance in detecting targets more accurately in complex environments and can effectively detect small targets, for example, long-distance pedestrians and signal lights. Ablation experiment results showed that each module contributed to the whole performance, and the mean average precision decreased when a module was removed.

Author Contributions: Conceptualization, Z.L. and W.Z.; methodology, W.Z.; formal analysis, Z.L., X.Y., and W.Z.; investigation, W.Z.; data curation, W.Z. and Z.L.; writing, Z.L. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: All data generated from this study are included in this published article. Raw data are available from the corresponding author on reasonable request.

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

References


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