electronics

Article

Occluded Vehicle Detection via Multi-Scale Hybrid Attention Mechanism in the Road Scene

Tianmin Deng *, Xuhui Liu and Li Wang

School of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China
* Correspondence: dtianmin@cqjtu.edu.cn

Abstract: The obstruction of vehicles by surrounding vehicles, obstacles, etc. is a common phenomenon in the practical application of automatic driving. In view of the problem that the vehicle's vision is affected by the occlusion, the vehicle feature information is incomplete, resulting in the low detection accuracy of the occlusion vehicle, and the occlusion vehicle detection method based on the multi-scale hybrid attention mechanism is proposed. The paper aims to fully excavate the advantages of multi-scale feature extraction, channel/space attention and other modules, and to design a multi-scale hybrid attention module suitable for occlusion vehicle detection to improve the detection accuracy of occlusion vehicles. Multi-scale features are enriched by the grouping convolution of different sizes of multi-scale feature extraction networks, and the parallel connection channels and spatial attention modules form different scale hybrid domain attention modules, which enhance the local feature information of the occluded vehicles and realize the reinforcement learning of multi-scale features and the suppression of occlusion interference information. Experimental results show that in the self-made occlusion vehicle dataset and the BDD100K occlusion vehicle dataset, the average mean accuracy of this method is 95.2% and 59.3%, respectively, which is 1.5% and 2.9% higher than that of the baseline network YOLOv5, respectively.

Keywords: occluded vehicle detection; multi-scale feature extraction; channel attention mechanism; spatial attention mechanism; hybrid domain attention module

Citation: Deng, T.; Liu, X.; Wang, L. Occluded Vehicle Detection via Multi-Scale Hybrid Attention Mechanism in the Road Scene. *Electronics* **2022**, *11*, 2709. https://doi.org/10.3390/electronics11172709

Academic Editor: Jose Eugenio Naranjo

Received: 6 July 2022
Accepted: 24 August 2022
Published: 29 August 2022

1. Introduction

Efficient and accurate vehicle detection is very important for intelligent transportation systems, and it is also an important task for driverless vehicles to perceive the road environment [1]. With the development of CNN and other deep learning technologies, vision-based target detection methods have achieved exciting results and have been widely used in many fields. Machine vision technology is also widely used in the perception module of intelligent transportation systems. However, the traffic environment of urban roads is complex and changeable, which brings great challenges to vehicle detection based on computer vision.

In the following scene of autonomous vehicles, there are often problems that vehicles block each other, or vehicles are blocked by background information. When there is mutual occlusion between vehicles, the occluded vehicles will lose some feature information due to local coverage, while the features of the non-occluded parts will easily bring some interference to the occluded vehicles, resulting in the impact of the detection results. When the vehicle is blocked by irrelevant background information, the blocking of irrelevant obstacles will also cause the loss of vehicle features, weaken the expression ability of the key feature information of the vehicle to be inspected, and make it difficult for the target detector to extract and learn its features. Therefore, the occlusion problem in complex scenes is similar to that of small targets and low illumination, which is one of the most challenging road vehicle detection tasks at present. In an environment with dense vehicles, the local obstruction of the vehicle running out in front or on the side by the obstacles may

cause the autonomous vehicle to be unable to find and avoid danger in time. Therefore, it is essential to quickly and accurately capture the key information of the vehicle in the occluded scene, which is helpful for autonomous vehicles to discover potential dangers and avoid them in the process of following the vehicle on urban roads.

Although there are many researches on vehicle detection [2,3], few of them consider occlusion. In road driving scenes, vehicle occlusion is very common. There are a total of 35.8% of the annotated vehicles and objects that are occluded on the KITTI dataset [4]. In order to make the target detector make full use of the local features of the occluded vehicle and minimize the interference caused by obstacles, a multi-scale hybrid attention module (MHAM) is designed to enhance the local feature information of the vehicle to be inspected and suppress the interference of irrelevant occlusion information. Combined with the above proposed MHAM strategy, this paper proposes a multi-scale hybrid attention module-you only look once (MHAM-YOLO) method based on multi-scale hybrid attention mechanism.

The remainder of this paper is organized as follows. Section 2 provides some related works of vehicle detection and occlusion detection, and offers the proposed multi-scale hybrid attention module-you only look once method. Section 3 shows the experiments and results of the proposed method. Section 4 presents our conclusion.

2. Related Works

Vehicle target detection. At present, vehicle target detection algorithms based on deep learning can be divided into two categories: a two-stage detection algorithm based on region recommendation and a one-stage detection algorithm based on regression.

The two-stage algorithm needs to form a preselection box, and then carry out fine-grained object detection. The detection accuracy is high, but the detection efficiency is low. The representative algorithms are: R-CNN, Fast R-CNN, Faster R-CNN and mask R-CNN [5]. Many vehicle detection methods based on improved R-CNN are proposed and achieve high detection accuracy [6,7]. However, these methods are relatively poor in real-time performance, which limits the application of these methods in actual driving scenes [8].

The single-stage algorithm does not need to generate a preselection box. Compared with the two-stage algorithm, the single-stage algorithm has faster detection speed. Representative algorithms include Retinanet [9], YOLO [10] and single shot multi-box detector (SSD) [11]. The improved YOLO is widely used for vehicle detection because of its good performance in efficiency and accuracy. However, many improved methods have no obvious effect on occluded vehicle detection.

Occlusion detection. At present, most researches on the recognition of occluded images mainly focus on large scale traditional objects of visible images, such as occluded pedestrian detection [12], occluded face detection [13], and so on. For occluded vehicle detection, most works focus on aerial images instead of the road scene [14,15]. A part-aware region based on Faster R-CNN is proposed for occluded vehicle detection in the road scene [4], but it achieves poor real-time. In view of the uncertainty of occlusion angle, layer degree and occlusion range, and the difficulty of occluded vehicle recognition, a multi-level optimization algorithm for occluded vehicle recognition is proposed [16]. On this basis, in order to make the target detector make full use of the local characteristics of the occluded vehicle and minimize the interference caused by obstacles, a multi-scale hybrid attention module is designed to improve the detection accuracy of the occluded vehicle.

2.1. Proposed Network Architecture

In order to improve the accuracy and efficiency of occluded vehicle recognition, an occluded vehicle detection model based on multi-scale hybrid attention mechanism is designed in this paper. The overall network structure is shown in Figure 1. In the figure, C1 to C5 are standard convolutional layers, extracting feature information in the image, P3 to P5 transmit deep semantic features from top to bottom, and N3 to N5 transmit target location information from bottom to top.

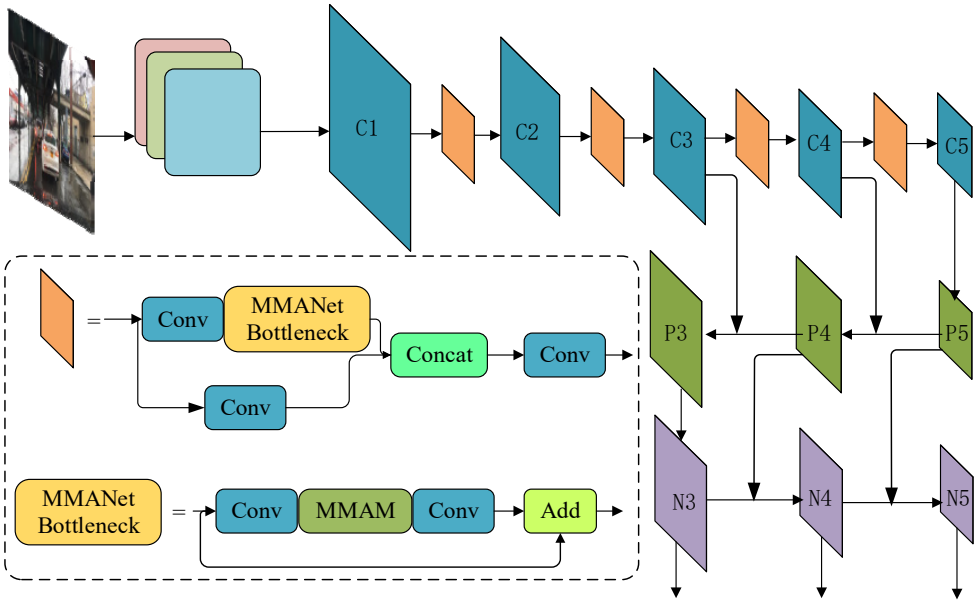


Figure 1. Architecture of proposed MHAM-YOLO network.

The MHAM-YOLO method uses a multi-scale hybrid attention module (MHAM) embedded in the bottleneck layer of the feature extraction network to enhance the local feature information of the vehicle to be inspected that is not occluded and suppress the interference of irrelevant occlusion information. MHAM includes three modules: multi-scale feature extraction, channel attention and spatial attention. The multi-scale feature extraction module is located in the front end to collect different scale features output from different receptive field convolution layers, and then use the channel and spatial attention modules on different scale feature layers to generate attention weights between channels and within feature map pixels, respectively. In the multi-scale features with large amounts of information, the important information that is more conducive to the occlusion vehicle detection task is filtered out.

2.2. Multi-Scale Feature Extraction Module

In the feature extraction stage of the network, if only the fixed scale convolution check input feature layer is used for feature extraction, the defect of a relatively single receptive field will make it difficult to capture rich context information, resulting in the extracted occluded vehicle features being limited to a single scale [17]. In order to make full use of the different receptive field information of the input feature space at multiple scales, this paper adopts the multi-scale grouping convolution method to mine the feature space from multiple scales, and a multi-scale feature extraction module (MFEM) is proposed and shown in Figure 2.

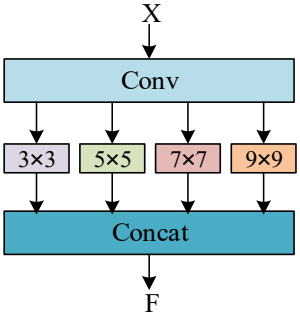


Figure 2. Multi-scale feature extraction module.

Suppose that the input feature space of MFEM can be expressed as $\boldsymbol{X} = [x_1, x_2, \cdots, x_c] \in \mathbf{R}^{C \times H \times W}$, then the split segmentation method is used to divide it evenly from the channel into n parts. If the channel size of the input feature space is C, the number of channels of each feature map after the segmentation operation can be expressed as $C' = C/\mathrm{n}$. In order to reduce the computational cost, the features of each feature map are collected by n groups:

$$\boldsymbol{F}_i = F_{\mathrm{conv}}^{k_i \times k_i}(\boldsymbol{X}_i, G_i), i = 0, 1, 2, \ldots, \mathrm{n}-1. \tag{1}$$

where the $F_{\mathrm{conv}}^{k_i, k_i}(\boldsymbol{X}_i, G_i)$ represents the convolution of $\boldsymbol{X}_i$ with kernel size $k_i \times k_i$ and groups number G_i. In order to further reduce the computational cost of the MFEM module, the value of n is set to 4, that is, the input feature space is divided into four feature maps with the same number of channels. The number of packets corresponding to the packet convolution of these four parts is set to 1, 2, 3 and 4, while the resolution of the packet convolution kernel is set to 3, 5, 7 and 9. The features of four scales are fused on the channel through the concat splicing operation:

$$\boldsymbol{F} = \mathrm{Concat}([F_0, F_1, F_2, \ldots, F_{\mathrm{n}-1}]). \tag{2}$$

The MFEM fuses the extracted multi-scale feature information on the channel, and the output feature space can carry multi-scale powerful context information. Therefore, MFEM effectively improves the negative impact of a single size convolution kernel on the feature expression ability of the target detector.

2.3. Hybrid Domain Attention Structure

There are two kinds of attention mechanisms for image feature extraction: channel attention and spatial attention.

Channel attention. In the target detector, the smaller the resolution of the characteristic image, the more the number of channels. When the number of layers of the model is deeper and deeper, the low-resolution feature map has a large amount of channel information. Therefore, it is often very difficult for the target detector to filter out the important channel information from a large number of channels. The channel attention mechanism can help the target detector to find more important channels for the target detection task and suppress those unimportant channels. Squeeze-and-excitation module (SEM) [18] is a representative channel attention mechanism, and its structure is shown in Figure 3.

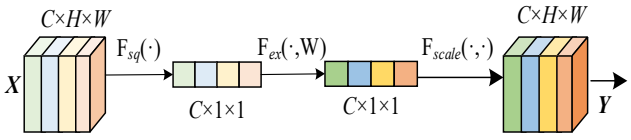


Figure 3. Squeeze-and-excitation module.

The attention mechanism of the SEM can be divided into three parts: squeeze, excitation and scale. First, a squeeze mapping is performed to squeeze the original feature space into an output feature space $\boldsymbol{Z} \in \mathbf{R}^{C \times H \times W}$:

$$z_c = F_{\mathrm{s}}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j). \tag{3}$$

where the $x_c \in \mathbf{R}^{H \times W}$ represents the feature map of input feature space $\boldsymbol{X}$ in the channel, $F_{\mathrm{s}}(\cdot)$ represents the squeeze mapping, H and W are the height and width of the feature map.

Secondly, the excitation is performed to appropriately reduce the required computational cost, and the corresponding channel attention weight coefficients are generated for the feature map on all channels. It could be expresses as:

$$\boldsymbol{S} = F_{\mathrm{e}}(\boldsymbol{Z}, \boldsymbol{W}) = \sigma(\boldsymbol{W}_1 \delta(\boldsymbol{W}_0 \boldsymbol{Z})). \tag{4}$$

where the $F_{\mathrm{e}}(\cdot, \boldsymbol{W})$ represents the excitation mapping, $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ represent the parameters of the first and second full connection layers, respectively, δ and σ are the ReLU function and sigmoid function, respectively.

Finally, the scale is performed to weight the $\boldsymbol{S}$ and $\boldsymbol{X}$, and the output of SEM is obtained in the feature space $\boldsymbol{Y} = [y_1, y_2, \cdots, y_c] \in \mathbf{R}^{C \times H \times W}$:

$$y_c = F_{\mathrm{scale}}(x_c, s_c) = x_c \times s_c. \tag{5}$$

where the F_{scale} represents the channel attention weighted mapping function.

Spatial attention. Different from the channel attention mechanism, the spatial attention mechanism assigns the same attention weight graph to each channel. The spatial attention module (SAM) [19] has the same resolution as the feature map and corresponds to each pixel on the feature map. Its structure is shown in Figure 4.

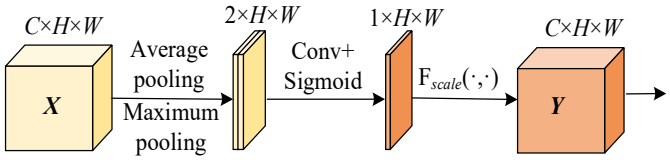


Figure 4. Spatial attention module.

The spatial attention module applies the attention weight map to the input feature to strengthen the features of key areas, and it could be expressed as:

$$\boldsymbol{M} = \sigma\left(F_{\mathrm{conv}}^{7\times7}(\mathrm{Concat}([\boldsymbol{X}_{\mathrm{avg}}, \boldsymbol{X}_{\mathrm{max}}]))\right). \tag{6}$$

where the $\boldsymbol{X}_{\mathrm{avg}} \in \mathbf{R}^{1 \times H \times W}$ and $\boldsymbol{X}_{\mathrm{max}} \in \mathbf{R}^{1 \times H \times W}$ represent the feature map after global average pooling and global maximum pooling, $F_{\mathrm{conv}}^{7\times7}$ represents the convolution with 7×7 kernel size, and σ is the sigmoid function.

By using the spatial attention weight map, the output of SAM could be expressed as:

$$\boldsymbol{Y} = F_{\mathrm{scale}}(\boldsymbol{X}, \boldsymbol{S}) = \boldsymbol{X} \otimes \boldsymbol{M} \tag{7}$$

where the F_{scale} represents the spatial attention weighted mapping function, $\otimes$ represents the weighted multiplication.

The SEM performs global attention weighting on the feature information in different channels of the input feature space, but it does not take into account the information interaction between the internal regions of the feature map. The SAM uses the generated attention weight graph to apply the same weight graph to each channel of the input feature space, while it ignores the information interaction between each channel.

In order to take advantage of the two attention mechanisms, some hybrid domain attention mechanisms are proposed [20,21]. There are two main mixing methods: a cascading and a parallel combination. The hybrid domain attention mechanism of the cascading combination connects two different dimensions of attention through cascading. The hybrid domain attention mechanism of the parallel combination combines the information of the channel and space through a parallel method. Compared with the cascade combination method, the parallel combination method does not need to consider the arrangement order of the two kinds of attention. The channel and spatial attention learn the key information of the original feature space at the same time, and only one step is needed to weight the

input feature space. Therefore, this paper adopts the parallel connected hybrid domain attention structure, and its structure is shown in Figure 5.

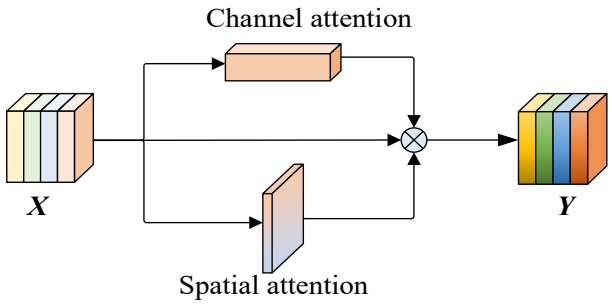


Figure 5. Hybrid domain attention structure for parallel combinations.

2.4. Multi-Scale Hybrid Attention Module

By using the MFEM and parallel connected hybrid domain attention structure, the MHAM was constructed, and it is shown as Figure 6.

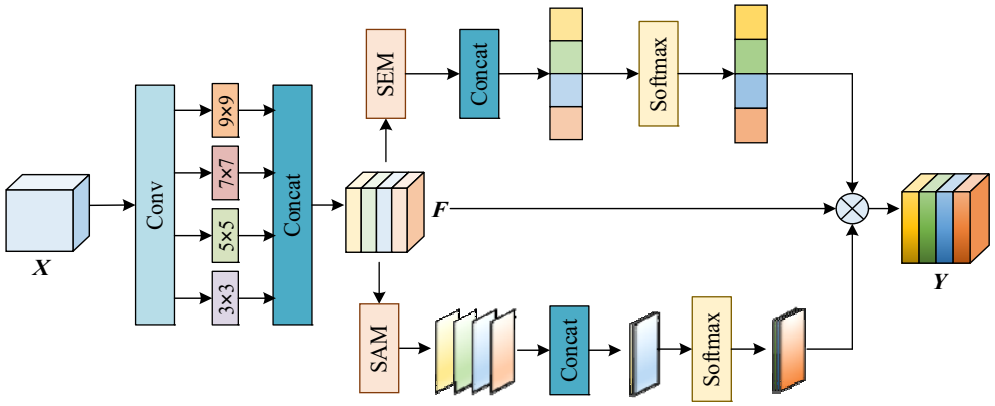


Figure 6. Multi-scale hybrid attention module.

According to Figure 6, the multi-scale features firstly were obtained by the MFEM and concatenate. Then, the SEM was used to learn the features of different scales, so as to generate the initial channel attention weight:

$$\boldsymbol{S}_i = \mathrm{SEM}(\boldsymbol{F}_i), i = 0, 1, 2, \ldots, \mathrm{n}-1 \tag{8}$$

$$\boldsymbol{S} = \mathrm{Concat}([\boldsymbol{S}_0, \boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_{\mathrm{n}-1}]) \tag{9}$$

In order to form a more stable long-term important relationship between multi-scale feature channels, Softmax function is used to perform recalibration:

$$\boldsymbol{cha}_i = \mathrm{softmax}(\boldsymbol{S}_i) = \frac{\exp(\boldsymbol{S}_i)}{\sum_{i=0}^{\mathrm{n}-1} \exp(\boldsymbol{S}_i)} \tag{10}$$

Similarly, by using the SAM, the weight map of the multiple scale feature maps in the initial spatial would be obtained. It could be expressed as:

$$\boldsymbol{M}_i = \mathrm{SAM}(\boldsymbol{F}_i), i = 0, 1, 2, \ldots, \mathrm{n}-1. \tag{11}$$

$$\boldsymbol{M} = \boldsymbol{M}_0 + \boldsymbol{M}_1 + \ldots + \boldsymbol{M}_{\mathrm{n}-1}. \tag{12}$$

In order to realize the information exchange within the multi-scale feature map, the Softmax function can also be used to perform the recalibration operation to obtain the multi-scale spatial attention weight map.

$$\boldsymbol{spa}_i = \text{softmax}(\boldsymbol{M}_i) = \frac{\exp(\boldsymbol{M}_i)}{\sum_{i=0}^{n-1} \exp(\boldsymbol{M}_i)} \tag{13}$$

Using the multi-scale channel attention weight $\boldsymbol{cha}_i$ and the spatial attention weight $\boldsymbol{spa}_i$, the weighted multi-scale feature space could be obtained:

$$\boldsymbol{Y}_i = \boldsymbol{F}_i \otimes \boldsymbol{cha}_i \otimes \boldsymbol{spa}_i, i = 0, 1, 2, \ldots, n-1. \tag{14}$$

Finally, by using the concat operation, the final output feature space of MHAM is constructed:

$$\boldsymbol{Y} = \text{Concat}([\boldsymbol{Y}_0, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{n-1}]). \tag{15}$$

In the case of focusing on the multi-scale features, MHAM gives the target detector the ability to mine key information and capture more effective information of occluding vehicles. In this paper, MHAM is embedded in the bottleneck layer of the target detector feature extraction network.

3. Experiment

3.1. Implementation Details of Experiment

3.1.1. Dataset and Annotations

In order to test the performance of the proposed network, the occluded vehicle dataset was made in the road scene. The video of road vehicles under various complex road conditions in Chongqing was collected by camera, and a total of 5123 image data containing occluded vehicles were screened, including 3500 training set images, 811 verification set images, and 812 test set images.

The vehicles in the picture were labelled and the corresponding label information for the training and testing of the model was generated. Figure 7 shows the visual schematic diagram of some samples of the dataset and their label information.

Figure 7. Sample of the homemade dataset section label.

All vehicles were divided into four categories: car, bus, truck and van. A total of 28859 annotation information was obtained, and the number of instances of different types of vehicles is shown as Table 1.

Table 1. Number of instances of different types of vehicles.

Vehicle Category	Car	Bus	Truck	Van
Number of instances	23,248	1686	1265	2660

3.1.2. Platform and Parameters

The experiment was conducted under the environment of I Intel i5-10400 2.90 GHz, and Nvidia RTX 3060.

In reference to the YOLOv5 network, the proposed MHAM-YOLO network was implemented by Pytorch framework in this paper. The Adam algorithm was used to optimize the model parameters. The weight decay was 5×10^{-4}, the initial learning rate was 1×10^{-2}, and the learning rate was decayed once in each round. The decay rate was 0.95. A total of 200 rounds were trained. There were 16 batches in the round, with 320 training samples in each batch.

In order to evaluate the effectiveness of the occluded vehicle detection algorithm proposed in this paper, the number of parameters and floating-point operations (FLOPs) per second were selected to evaluate the complexity of the model, and the mean average precision (mAP) was selected as the evaluation index for the comprehensive detection performance of the model for multiple target categories, The average precision (AP) was used to evaluate the detection performance of the model for a single target category.

3.1.3. Model Training

The YOLOv5 network was employed as the baseline algorithm. Both the YOLOv5 and the proposed MHAM-YOLO network were trained on self-made occluded vehicle dataset, and the average loss convergence curve during training is shown as Figure 8.

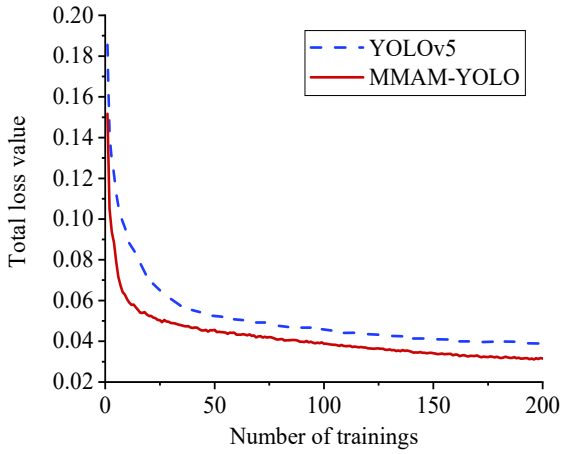
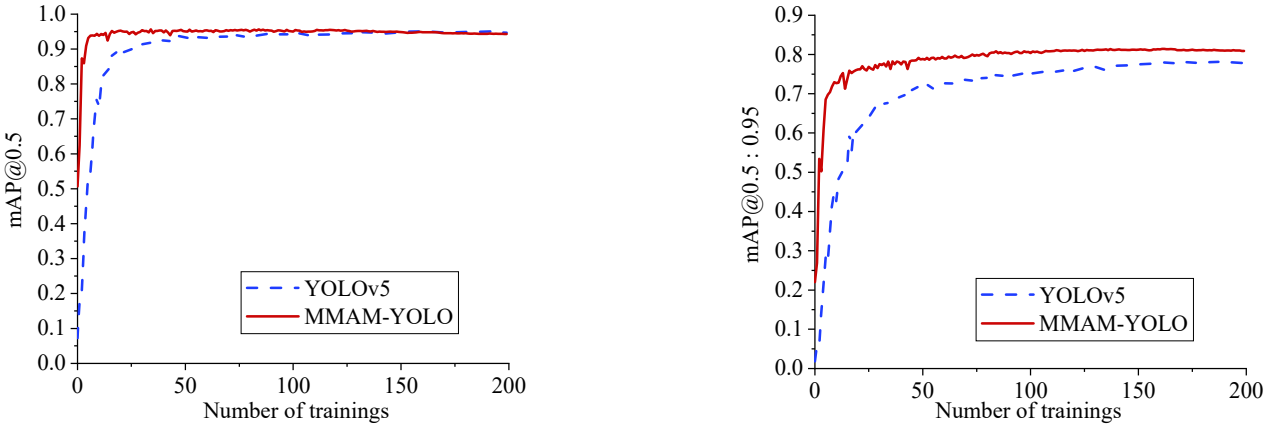


Figure 8. Total loss convergence curve.

As Figure 8 shows, the proposed MHAM-YOLO could decrease faster than that of YOLOv5 when the training times were less than 50. This indicates that the model in this paper has a faster learning speed for occluded vehicle features.

Moreover, the mAP@0.5 and mAP@0.5: 0.95 of the baseline algorithm and MHAM-YOLO method were compared, as shown in Figure 9.

The comparison in Figure 9a shows that the mAP@0.5 curve of MHAM-YOLO immediately tended to be stable after a rapid rise. Then it began to decline slightly when it was trained to about 120 times, and it gradually fell below the baseline algorithm after the 160th training. However, Figure 9b shows that the mAP@0.5: 0.95 curve of MHAM-YOLO has great advantages.



(**a**) mAP@0.5 (**b**) mAP@0.5: 0.95

Figure 9. Comparison of training accuracy curves.

3.2. Experiment and Analysis

3.2.1. Ablation Experiment

In order to verify the effectiveness of the multi-scale hybrid attention mechanism proposed in this paper, the channel attention module SEM, spatial attention module SAM, hybrid domain attention module and multi-scale hybrid attention module MHAM were verified and analyzed on the self-made occluded vehicle dataset. The test results are shown in Table 2.

Table 2. Ablation study of ASPP and FPN blocks.

Method	Param. (M)	FLOPs (G)	FPS	AP/%				mAP/%
				Car	Bus	Truck	Van	
YOLOv5	46.65	114.3	38	97.2	92.8	91.7	93.1	93.7
YOLOv5 + SEM	48.43	118.7	37	97.8	93.1	92.0	93.9	94.2
YOLOv5 + SAM	46.87	114.9	38	97.4	92.8	91.6	93.9	93.9
YOLOv5 + SEM + SAM	48.65	119.4	37	97.9	93.2	92.1	94.9	94.5
MHAM-YOLO	54.18	140.8	35	98.8	93.7	92.5	95.6	95.2

The ablation experiment results show that after the SEM was implanted in the feature extraction network of the baseline algorithm, the number of model parameters increased to 48.43 M, the number of floating-point operations increased to 118.7 G, and the model detection rate decreased by 1 frames/s. At the same time, good detection accuracy was achieved. The mAP value was increased from 93.7% of the original model to 94.2%, and the AP values of the four vehicle categories were also improved to some extent. It can be seen that the remote dependency between channels modeled by SEM was very effective for the detection of occluded vehicle targets.

After SAM was embedded in the baseline algorithm, the map of the model was improved by 0.2%, and only 0.6 G of floating-point operations and 0.22 M of parameters were additionally increased. Compared with SEM, the mAP obtained after SAM was embedded in the baseline algorithm was lower, but it also ensured a small model scale and parameter quantity. Therefore, it was also effective for capturing the feature information of key areas from the spatial dimension using Sam.

After embedding the parallel hybrid domain attention module combined with SEM and SAM into the feature extraction network of the baseline algorithm, the number of model parameters and floating-point operations increased by 2 M and 5.1 G, respectively, and reached 94.5% of mAP. It can be seen that the hybrid domain attention effectively combined SEM and SAM in a parallel way, and fully played the role of screening effective

feature information. After end-to-end learning in the training phase, the generated attention weight was very effective, and could effectively improve the detection performance of occluded vehicles while ensuring small model complexity.

On the basis of the hybrid domain attention module, MFEM was introduced as the source of multi-scale feature space to obtain the multi-scale hybrid attention module MHAM. Then the MHAM was integrated into the bottleneck layer of the baseline algorithm to obtain the MHAM-YOLO method. The experimental results showed that with the embedding of MHAM, the parameters of the model increased to 54.18 M, and the number of floating-point operations also increased to 140.8 G, The detection rate of the model was reduced from 38 frames/s of the baseline algorithm to 35 frames/s, but it could still meet the requirements of real-time detection. At the same time, the mAP value of MHAM-YOLO reached 95.2%, which is 0.7% higher than that of the model only embedded with the hybrid domain attention module, and 1.5% higher than that of the baseline algorithm. The AP values of the four types of occluded vehicles were also improved to varying degrees. Therefore, the method of combining hybrid domain attention with multi-scale feature extraction module was very effective. Hybrid domain attention could fully extract multi-scale features from the multi-scale feature extraction module, and then mine the key feature information on multiple scales for shelter vehicle detection.

To sum up, under the condition of meeting the real-time performance, MHAM-YOLO algorithm first extracts the feature space information of different scales through the multi-scale feature extraction module, and then sends the extracted multi-scale feature information to the hybrid domain attention module composed of the spatial attention module and the channel attention module to mine the multi-scale information of the feature space accurately, so as to suppress the feature information of irrelevant occlusion, strengthen the feature information of the non-occluded part of the occluded vehicle, and effectively perform the road vehicle detection task in the occluded scene.

3.2.2. Comparison of Different Attention Modules

In order to further evaluate the detection performance of the multi-scale hybrid attention mechanism proposed in this paper, the multi-scale hybrid attention module MHAM was compared with the existing advanced hybrid domain attention mechanisms BAM [22] and CBAM [23] on the self-made occluded vehicle dataset. The experimental results are shown in Table 3. In this experiment, YOLOv5 was still used as the baseline algorithm, and BAM was embedded into the back layer of each C3 layer in the baseline algorithm feature extraction network for training. The embedding method of CBAM was the same as that of BAM. Then, the models of these two hybrid domain attention mechanisms were compared with the baseline algorithm and the MHAM-YOLO in this paper.

Table 3. Comparison of different attention modules.

	Method	Param. (M)	FLOPs (G)	FPS	AP/%				mAP/%
					Car	Bus	Truck	Van	
A	YOLOv5	46.65	114.3	38	97.2	92.8	91.7	93.1	93.7
B	BAM-YOLO	47.31	116.0	38	97.4	92.5	90.3	94.5	93.7
C	CBAM-YOLO	48.45	118.9	37	98.2	92.6	91.9	94.1	94.2
D	MHAM-YOLO	54.18	140.8	35	98.8	93.7	92.5	95.6	95.2

It can be seen from the comparative results in Table 3 that the performance of the model has not been substantially improved after BAM was embedded in the baseline algorithm. Not only did the mAP value still maintain at 93.7% of the baseline algorithm, but it also brought a certain number of parameters and floating-point operations. From the AP values of various vehicle categories, we found that the AP value of BAM for van category was greatly improved, which was 1.4% higher than that of the baseline algorithm, and the AP values of bus and truck categories were lower than that of the baseline algorithm. Therefore,

BAM did not achieve ideal results in the occluded vehicle scene in this paper. Further experiments show that after the baseline algorithm was embedded in CBAM, the number of model parameters and floating-point operations were increased to 48.45 M and 118.9 G, respectively, and the model detection rate was slightly reduced to 37 frames/s. It was evident that the embedding of CBAM did not bring too much computational overhead to the model. From the change of map value, we found that CBAM achieved 94.2% of mAP, and at the same time, 1% of AP values of car and van were significantly improved. It could effectively enhance the detection performance of baseline algorithm. Compared with the above two hybrid domain attention mechanisms, the multi-scale hybrid attention module in this paper achieved the best detection performance. The BAM embedding failed to bring good detection results to the baseline algorithm, and was not an effective method for detecting occluded vehicles. Compared with CBAM, the MHAM method in this paper had 5.73 M higher parameters and 21.9 G higher floating-point operations, but only reduced the model detection rate of 2 frames/s. At the same time, the mAP value of the model was 1% higher than that of CBAM, and the AP value of each vehicle category had also achieved the optimal results. Therefore, the detection performance of MHAM for occluded vehicles was commendable compared with the current advanced hybrid domain attention mechanism.

3.2.3. Comparison of BDD100K Dataset

In order to further verify the effectiveness of this method in other occluded scenes, the BDD100K dataset was used for generalization experiments. The BDD100K dataset contained a large number of difficult samples such as night and fuzzy, and the detection accuracy was low. Here, 9904 pictures of cars, buses and trucks with "occlusion" label information were randomly selected from the BDD100K dataset, of which 8021 were used as the training set, 892 as the verification set and 991 as the test set, forming a generalized experimental dataset. The experimental results are shown in Table 4.

Table 4. Comparison of BDD100K dataset.

Method	Param. (M)	FLOPs (G)	FPS	AP/%			mAP/%
				Car	Bus	Truck	
YOLOv5	46.64	114.3	38	74.2	42.4	52.7	56.4
MHAM-YOLO	54.17	140.8	35	74.5	47.9	55.4	59.3

Table 4 shows that on the BDD100K occluded vehicle dataset, the parameter quantity of MHAM-YOLO algorithm increased by 7.53 M and the number of floating-point operations by 26.5 G compared with the baseline algorithm, but only the model detection rate of 3 frames/s was reduced, and the average accuracy of each category was improved. Among them, the average accuracy of cars, buses and trucks increased by 0.3%, 5.5% and 2.7%, respectively, and the mAP increased by 2.9%, This fully proves that the MHAM-YOLO algorithm in this paper can detect occluded vehicles better.

3.3. Visualization of Occluded Vehicle Detection

In order to more intuitively evaluate the detection effect of the proposed MHAM-YOLO algorithm on vehicle targets in occluded scenes, this paper tests the actual detection effect of the baseline algorithm YOLO5 and the MHAM-YOLO algorithm on the self-made occluded vehicle dataset. This paper selects some sample images of occluded vehicles from the detection results of the two algorithms for comparative analysis, as shown in Figure 10. Firstly, the detection effects of the two algorithms in the case of serious vehicle occlusion are compared.

(**a**) results of YOLOv5 (**b**) results of MHAM-YOLO

Figure 10. Comparison of the detection effect of vehicle obstruction.

From the comparison between Figures 10a and 10b, it can be seen that the vehicle in front of the road is seriously blocked by the vehicle behind it, which leads to the problem of repeated detection when the baseline algorithm detects the truck. However, the MHAM-YOLO algorithm in this paper shows strong anti-interference ability when facing the situation of mutual occlusion of vehicles. By weakening the interference of blocked vehicles and strengthening the characteristics of blocked vehicles, the vehicle target is accurately detected. Subsequently, this paper compares the dense continuous occlusion between vehicles, and the results are shown in Figure 11.

(**a**) results of YOLOv5 (**b**) results of MHAM-YOLO

Figure 11. Comparison of the detection effect of dense continuous occlusion of vehicles.

By further comparing Figure 11a with Figure 11b, we found that even in the face of a more complex scene of dense and continuous occlusion of vehicles, the method in this paper can also achieve a very ideal detection effect. It can be seen that while the baseline algorithm fails to detect two vehicles blocked by surrounding vehicles, and there are also false detections, MHAM-YOLO relies on its good ability to detect occluded vehicles to effectively detect vehicles that have been occluded with most of the key features. Finally, this paper compares the ability of the two methods to deal with background occlusion, as shown in Figure 12.

From the comparison between Figures 12a and 12b, it can be seen that the baseline algorithm has a serious impact on the feature extraction of the blocked vehicles due to the blocking of the roadside guardrail on the vehicles in front. Therefore, there is a missed detection situation. While MHAM-YOLO fully suppresses the interference of the roadside guardrail and other complex backgrounds through the multi-scale hybrid attention module, only a small part of the feature information that is not occluded is used to complete the accurate detection of cars and trucks.

(**a**) results of YOLOv5 (**b**) results of MHAM-YOLO

Figure 12. Comparison of detection effects of irrelevant background information occlusion.

The above analysis shows that the proposed road vehicle detection method MHAM-YOLO in occluded scenes can use the proposed multi-scale hybrid attention mechanism to well complete the task of detecting occluded vehicles. It can fully mine the important features of occluded vehicles by using the screening function of the attention mechanism, whether in the case of mutual occlusion between vehicles or the case of background information on vehicles. Hence, it can provide guidance for the detection process of target detector.

4. Conclusions

An occluded vehicle detection method MHAM-YOLO based on multi-scale hybrid attention mechanism is proposed. A multi-scale hybrid attention module (MHAM) was designed, which integrates a multi-scale feature extraction module, a channel attention module and a spatial attention module. Firstly, the multi-scale feature extraction module will segment the original input features along the channel dimension and use the grouping convolution of different sizes to collect rich multi-scale features. Further, in the training phase of the target detector, the multi-scale features were learned through the parallel combined hybrid domain attention mechanism, and the hybrid domain attention weights on different scales were obtained to enhance the local feature information of the occluded vehicle. The self-made occluded vehicle dataset and the occluded vehicle dataset of BDD100K were used to evaluate the effectiveness of MHAM-YOLO. The experimental results show that the MHAM-YOLO algorithm performs better than the baseline algorithm in the vehicle detection task in an occluded environment. It can accurately detect whether the vehicle is occluded by the vehicle, or the vehicle is occluded by irrelevant background information.

Author Contributions: Conceptualization, T.D.; methodology, T.D.; validation, X.L. and L.W.; formal analysis, X.L. and L.W.; data curation, X.L.; writing—original draft preparation, T.D.; writing—review and editing, T.D. and X.L.; visualization, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [the National Key Research and Development Program of China] grant number [SQ2020YFF0418521], [Chongqing Science and Technology Development Foundation] grant number [cstc2020jscx-dxwtBX0019 and cstc2021jscx-gksbX0058] and [the Joint Key Research and Development Program of Sichuan and Chongqing (CN)] Grant number: [cstc2020jscx-cylhX0007 and cstc2020jscx-cylhX0005].

Data Availability Statement: The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments: This work was supported by the National Key Research and Development Program of China (Grant number: SQ2020YFF0418521), Chongqing Science and Technology Development Foundation (Grant number: cstc2020jscx-dxwtBX0019 and cstc2021jscx- gksbX0058), the Joint Key Research and Development Program of Sichuan and Chongqing (CN) (Grant number: cstc2020jscx-cylhX0007 and cstc2020jscx-cylhX0005). It is worth mentioning that all authors have contributed to the completion of the paper. The whole research process was led by Tianmin Deng

and assisted by Xuhui Liu and Li Wang. First, based on previous work, Tianmin Deng designed and completed the experiment, and then Li Wang and Xuhui Liu collected data and analyzed them. Finally, Tianmin Deng completed the writing of the paper.

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

References

1. Yang, Z.; Pun-Cheng, L.S.C. Vehicle detection in intelligent transportation systems and its applications under varying environments: A review. *Image Vis. Comput.* **2018**, *69*, 143–154. [CrossRef]
2. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Vehicle detection from UAV imagery with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
3. Wang, Z.; Zhan, J.; Duan, C.; Guan, X.; Lu, P.; Yang, K. A review of vehicle detection techniques for intelligent vehicles. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
4. Zhang, W.; Zheng, Y.; Gao, Q.; Mi, Z. Part-aware region proposal for vehicle detection in high occlusion environment. *IEEE Access* **2019**, *7*, 100383–100393. [CrossRef]
5. Bharati, P.; Pramanik, A. Deep learning techniques—R-CNN to mask R-CNN: A survey. *Comput. Intell. Pattern Recognit.* **2020**, *999*, 657–668.
6. Yang, W.; Li, Z.; Wang, C.; Li, J. A multi-task Faster R-CNN method for 3D vehicle detection based on a single image. *Appl. Soft Comput.* **2020**, *95*, 106533. [CrossRef]
7. Luo, J.Q.; Fang, H.S.; Shao, F.M.; Zhong, Y.; Hua, X. Multi-scale traffic vehicle detection based on faster R–CNN with NAS optimization and feature enrichment. *Def. Technol.* **2021**, *17*, 1542–1554. [CrossRef]
8. Yin, G.; Yu, M.; Wang, M.; Hu, Y.; Zhang, Y. Research on highway vehicle detection based on faster R-CNN and domain adaptation. *Appl. Intell.* **2022**, *52*, 3483–3498. [CrossRef]
9. Zhang, L.; Wang, H.; Wang, X.; Chen, S.; Wang, H.; Zheng, K. Vehicle object detection based on improved retinanet. *J. Phys. Conf. Series. IOP Publ.* **2021**, *1757*, 012070. [CrossRef]
10. Zhu, J.; Li, X.; Jin, P.; Xu, Q.; Sun, Z.; Song, X. Mme-yolo: Multi-sensor multi-level enhanced yolo for robust vehicle detection in traffic surveillance. *Sensors* **2020**, *21*, 27. [CrossRef] [PubMed]
11. Zhao, M.; Zhong, Y.; Sun, D.; Chen, Y. Accurate and efficient vehicle detection framework based on SSD algorithm. *IET Image Processing* **2021**, *15*, 3094–3104. [CrossRef]
12. Zhou, C.; Yuan, J. Occlusion pattern discovery for object detection and occlusion reasoning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2067–2080. [CrossRef]
13. Kumar, A.; Kumar, M.; Kaur, A. Face detection in still images under occlusion and non-uniform illumination. *Multimed. Tools Appl.* **2021**, *80*, 14565–14590. [CrossRef]
14. Zhang, W.; Liu, C.; Chang, F.; Song, Y. Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images. *Remote Sens.* **2020**, *12*, 1760. [CrossRef]
15. Du, S.; Zhang, P.; Zhang, B.; Xu, H. Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4. *IEEE Access* **2021**, *9*, 25671–25680. [CrossRef]
16. Zhu, J. *Research on Vehicle Recognition and Tracking Technology under Roadside Occlusion*; Southeast University: Nanjing, China, 2021.
17. Mao, G.-T.; Deng, T.-M.; Yu, N.J. UAV Aerial Image Object Detection Algorithm Based on Multi-scale Segmentation of Attention. *J. Aeronaut.* **2022**, *43*, 1–12. [CrossRef]
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
19. Jin, L.; Shu, X.; Li, K.; Li, Z.; Qi, G.J.; Tang, J. Deep ordinal hashing with spatial attention. *IEEE Trans. Image Processing* **2018**, *28*, 2173–2186. [CrossRef] [PubMed]
20. Roy, A.G.; Navab, N.; Wachinger, C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Trans. Med. Imaging* **2018**, *38*, 540–549. [CrossRef] [PubMed]
21. Wang, L.; Peng, J.; Sun, W. Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 884. [CrossRef]
22. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.