

Article

Occlusion Vehicle Target Recognition Method Based on Component Model

Haorui Han and Hanshan Li * 

School of Electronic and Information Engineering, Xi'an Technological University, Xi'an 710021, China; 15333687324@163.com

* Correspondence: lihanshan269@163.com

Abstract: As an important part of intelligent traffic, vehicle recognition plays an irreplaceable role in traffic management. Due to the complexity and occlusion of various objects in the traffic scene, the accuracy of vehicle target recognition is poor. Therefore, based on the distribution features of vehicle components, this paper proposes a two-stage VSRS-VCFM net occlusion vehicle target recognition method. Based on the U-net codec structure, combining multi-scale detection and double constraints loss to improve the visual region segmentation under complex background (VSRS) performance. At the same time, to establish the vehicle component feature mask (VCFM) module, based on the Swin Transformer backbone unit, combined with the component perception enhancement unit and the efficient attention unit, the extraction of the low-contrast component area of the vehicle target and the filtering of the irrelevant area are realized. Then, the component mask recognition unit is introduced to remove the occlusion component feature area and realize the accurate recognition of the occluded vehicle. By labeling the public data set and the collected data set, six types of vehicle component data sets are constructed for training, as well as design ablation experiments and comparison experiments to verify the trained network, which prove the superiority of the recognition algorithm. The experimental results show that the proposed recognition method effectively solves the problem of misrecognition and missing recognition caused by interference and occlusion in vehicle recognition.

Keywords: vehicle components; deep learning; components segmentation; Swin transformer; feature extraction; vehicle recognition



Citation: Han, H.; Li, H. Occlusion Vehicle Target Recognition Method Based on Component Model. *Appl. Sci.* **2024**, *14*, 11076. <https://doi.org/10.3390/app142311076>

Academic Editor: Andrea Prati

Received: 25 October 2024

Revised: 26 November 2024

Accepted: 26 November 2024

Published: 28 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

At present, the number of urban vehicles is increasing, and traffic congestion and traffic accidents are frequent, which brings difficulties to road monitoring and traffic supervision and has become an important factor hindering current social development [1]. In order to solve the traffic problem, all countries are actively studying intelligent transportation systems, and one of the important research contents of intelligent transportation systems is vehicle recognition [2,3]. Although computer vision recognition technology has made great progress, in the complex urban road environment, due to illumination, imaging deformation, vehicle target occlusion [4], and other factors, the camera often cannot capture the vehicle target image with high contrast and constant geometry, especially the complete vehicle target image, resulting in the loss of vehicle target information and the introduction of noise information. This results in a large intra-class difference in vehicle features, resulting in a great reduction in the accuracy of the recognition algorithm, which brings difficulties to the monitoring of traffic vehicles. Therefore, how to effectively recognize and analyze partially occluded vehicle targets is the focus of current research [5].

According to the relationship between the occlusion and the occluded object, occlusion can be divided into inter-target occlusion, perspective occlusion, and non-target occlusion [6]. Mutual occlusion between targets, such as mutual occlusion between vehicles, can

be solved by optimizing the detection method when the training samples are sufficient. The other is the occlusion of non-target objects. Due to the variety of occluded objects, there is often a lack of effective labeling information, which makes it difficult for the target detection module to process directly. At present, for the problem of inter-target occlusion and perspective occlusion, the overall analysis of the feature map is proposed to achieve the overall detection of the target, including the DeepParts [7] module and OR-CNN (occlusion-aware R-CNN) network [8]. These detection algorithms are based on partial semantics. Aiming at the occlusion problem caused by non-target objects [9], Devries and Taylor [10] proposed a regularization technique to enhance the data, called “Cutout”, to prevent the convolutional neural network from overfitting and causing poor occlusion adaptability. Cutout is similar to Dropout [11], which discards some of the features to enhance the robustness of the neural network. Cutout and Dropout delete features by covering black pixels or random noise to remove pixel information on the training image, but it will lead to information loss and low efficiency in training. Yun et al. [12] proposed a regularization method called “CutMix”. The strategy is to cut and paste occlusion blocks between the training images for image enhancement and blend the target area into the merged image proportionally, thus avoiding the negative impact of the network on the pixels that lack information in the training.

In recent years, the research on vehicle recognition can be divided into two categories [13]. The first is to restore the foreground occlusion part and realize the recognition method. Related methods can be roughly divided into autoencoder-based image restoration [14], a vehicle target detection algorithm based on the generative adversarial network [15], and a vehicle image detection method based on context-aware semantics [16]. Zhou et al. [17] proposed a multi-label learning approach to co-learn component detectors. For each component detector, the overall image of the component is used as training data. This joint learning method effectively improves the detection accuracy of the component detectors, but it cannot get rid of the large number of computational resources required to train the component detectors separately. Ren et al. [18] proposed the Recurrent Rolling Convolution (RRC) structure to improve the detection effect of the SSD algorithm on occlusion objects or small objects. However, this recognition method can only identify the occlusion generated between vehicles, and other objects are not good at identifying the occlusion of vehicles.

The other is based on the structural features of the occluded target, using the prior knowledge and the structural information of the visible part of the target to design the detector to improve the occlusion detection performance. The authors of [19] proposed a vehicle recognition method based on multi-scale feature fusion. It combines high-level and low-level vehicle image features and can effectively deal with the complex and changeable natural environment in which the vehicle is located. In [20], a deformable part-based model (DPM) is proposed to detect vehicles. That is, a model is trained for each type of vehicle, and then these models are used in the traffic scene image for vehicle detection and extraction and feature alignment. Finally, a support vector machine is used for classification and recognition, but a large number of data sets are needed for training. The authors of [21] propose a vehicle detection framework. Firstly, an accurate vehicle proposal network (AVPN) based on a hyper-feature map is used to generate candidate vehicle regions. Then, a vehicle is used to attribute the learning network (VALN) to verify and classify candidate regions.

In summary, researchers have conducted some research on occlusion object recognition to improve the reliability of occlusion detection. However, there is still inaccuracy in occlusion detection under the condition of large occlusion areas, complex backgrounds, and illumination change. Based on the advantages of deep learning in target recognition, considering the saliency and scale of vehicle targets, a vehicle target recognition network with component feature extraction is designed to improve its detection performance in occluded scenes. The contributions are summarized as follows:

- (1) In order to solve the target detection in complex backgrounds, the improved U-Net module is proposed to segment the visual saliency region and filter out the background interference.
- (2) The Swin Transformer is adopted as the main unit of component feature extraction to add a component perception enhancement unit and an efficient attention unit to deal with the impact of target vehicle scale change, weak contrast, and regional interference on component regional feature extraction so that the module can efficiently extract component features for component classification.
- (3) By establishing a CMR (Component Mask Recognition) unit is used to gather the multi-scale features from the occluded vehicle components to learn and generate the feature mask to realize the classification to recognize the occluded vehicle targets.

This paper includes the following outline. Section 2 introduces the structure of the VSRS-VCFM net. Section 3 introduces the experiments and results. The conclusions of this study are shown in Section 4.

2. Occlusion Vehicle Target Segmentation and Recognition Scheme Design

Aiming at the problem of vehicle target recognition under complex backgrounds, illumination change, and occlusion in urban traffic environments, the visual saliency mechanism is introduced to filter out the complex background. The Swin Transformer, receptive field expansion, and attention mechanism are used to reliably extract the vehicle target components. The Mask Decoder component is introduced to recognize the occluded vehicle in the component area distribution. The designed two-stage net VSRS-VCFM structure of occluded vehicle recognition is shown in Figure 1.

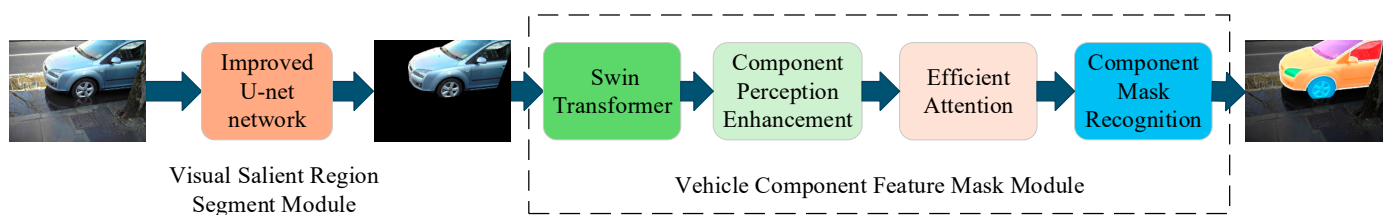


Figure 1. Vehicle recognition method based on component module.

The VSRS-VCFM (Visual Saliency Region Segment–Vehicle Component Feature Mask) network recognition method is mainly divided into two modules. First, in the Visual Saliency Region Segment Module, the improved U-NET is used to detect the visual saliency region of the detected image. Foreground objects, such as vehicles from the image, are segmented. The Vehicle Component Feature Mask module consists of four parts: the Swin Transformer unit, component perception enhancement unit, efficient attention unit, and CMR Unit. These modules can be used to extract the target low-contrast parts of the vehicle, filter out irrelevant areas, remove the feature areas of damaged parts, and provide effective data for classification. Vehicle identification is realized according to component distribution characteristics.

2.1. Design of Visual Saliency Region Segmentation (VSRS) Module Based on Complex Backgrounds

In order to realize the fast segmentation of the salient vehicle area in the image, the attention mechanism mimics human vision to find prominent areas that may be targets and then confirm them [22]. The U-Net [23] can be applied to saliency region detection, but when the vehicle object is in a cluttered background, the detection accuracy and speed will be affected. Aiming at this problem, a saliency target detection module is designed. The nested U-shaped structure can effectively extract the multi-scale features of each stage and aggregate the multi-stage multi-level features so as to improve the accuracy of vehicle target segmentation in complex backgrounds. The module structure is shown in Figure 2a. Each blue square is a Residual U-block (RSU). The downward line represents the down-sampling, the upward line represents the up-sampling, and the right line represents the

jump connection, that is, two features of the same height and width are connected in series on the channel dimension.

The structure of the RSU module is shown in Figure 2b. The structure includes up-sampling and down-sampling. Each layer of the net contains a convolution layer, a batch normalization layer, and an activation layer. H , W , and C represent the height, width, and number of channels of input and output; M is the measure of the number of input and output channels of each layer; the convolution kernel size of the convolution layer is 3×3 ; and D is the expansion rate of the dilated convolution.

The RSU is composed of a mixture of different receptive fields, so the VSRS module can obtain multi-scale target information and filter out local complex backgrounds. At the same time, the pooling operation of the RSU can improve the depth of the entire module architecture, facilitate the acquisition of global information, and filter out the global complex background. Since the VSRS module is based on the RSU and does not use any pre-trained backbone network for image classification, the constructed saliency target detection net module has flexible performance and strong adaptability, which is conducive to the realization of complex background filtering.

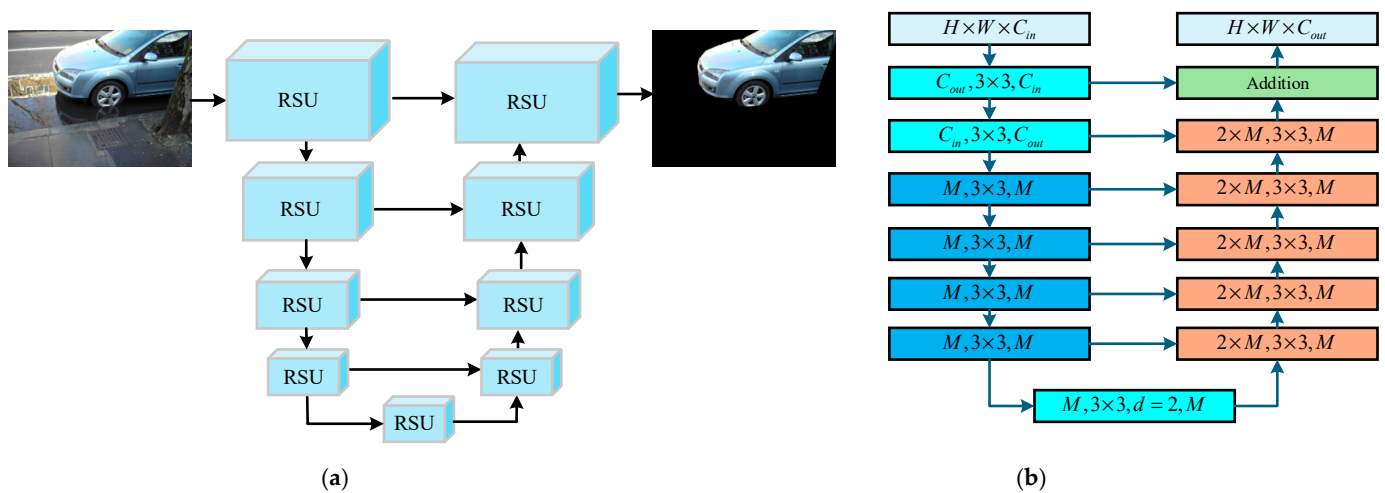


Figure 2. Visual saliency region segmentation module. (a) Visual saliency region segmentation structure, (b) residual U-block.

In the training process, the pre-training and supervised learning methods are used to optimize the net parameters. In order to consider both the module classification performance and the structure of the segmentation target, the loss function is composed of two parts: binary cross entropy loss and structural similarity loss. The binary cross entropy $loss_{BCE}$ is

$$loss_{BCE} = - \sum_{(x,y)}^{(H,W)} [L \cdot p(x,y) + (1-L) \log(1-p(x,y))] \quad (1)$$

In Equation (1), $p(x,y)$ and L represent the module's predicted value and the true value of the image coordinate point (x,y) with height and width W .

The second part of the loss function uses regional data to calculate the structural similarity loss as follows:

$$loss_{SSIM} = 1 - \frac{(2\mu_I\mu_S + C_1)(2\sigma_{IS} + C_2)}{(\mu_I^2 + \mu_S^2 + C_1)(\sigma_I^2 + \sigma_S^2 + C_2)} \quad (2)$$

In Equation (2), I and S represent the two images of the input and the output in the VSRS module to be compared, μ_I and μ_S are their average values, σ_I^2 and σ_S^2 are their variances, σ_{IS} is the covariance, and C_1 and C_2 are constants to increase stability.

2.2. Design of Vehicle Component Feature Mask Module Based on Swin Transformer

After obtaining the region of interest, it will not only segment the vehicle target but also segment other significant objects located in the foreground. It will interfere with the positioning of the vehicle target and need to filter it out. Considering that the features of the object components have better stability than the traditional corner points, colors, and other features, the regional features of the vehicle target components are selected for extraction. The components [24] in a vehicle target are defined as parts with different contour structures. How to effectively extract these parts of the vehicle is the key to subsequent vehicle target recognition. Therefore, a Vehicle Component Feature Mask (VCFM) module based on deep learning is designed, as shown in Figure 3.

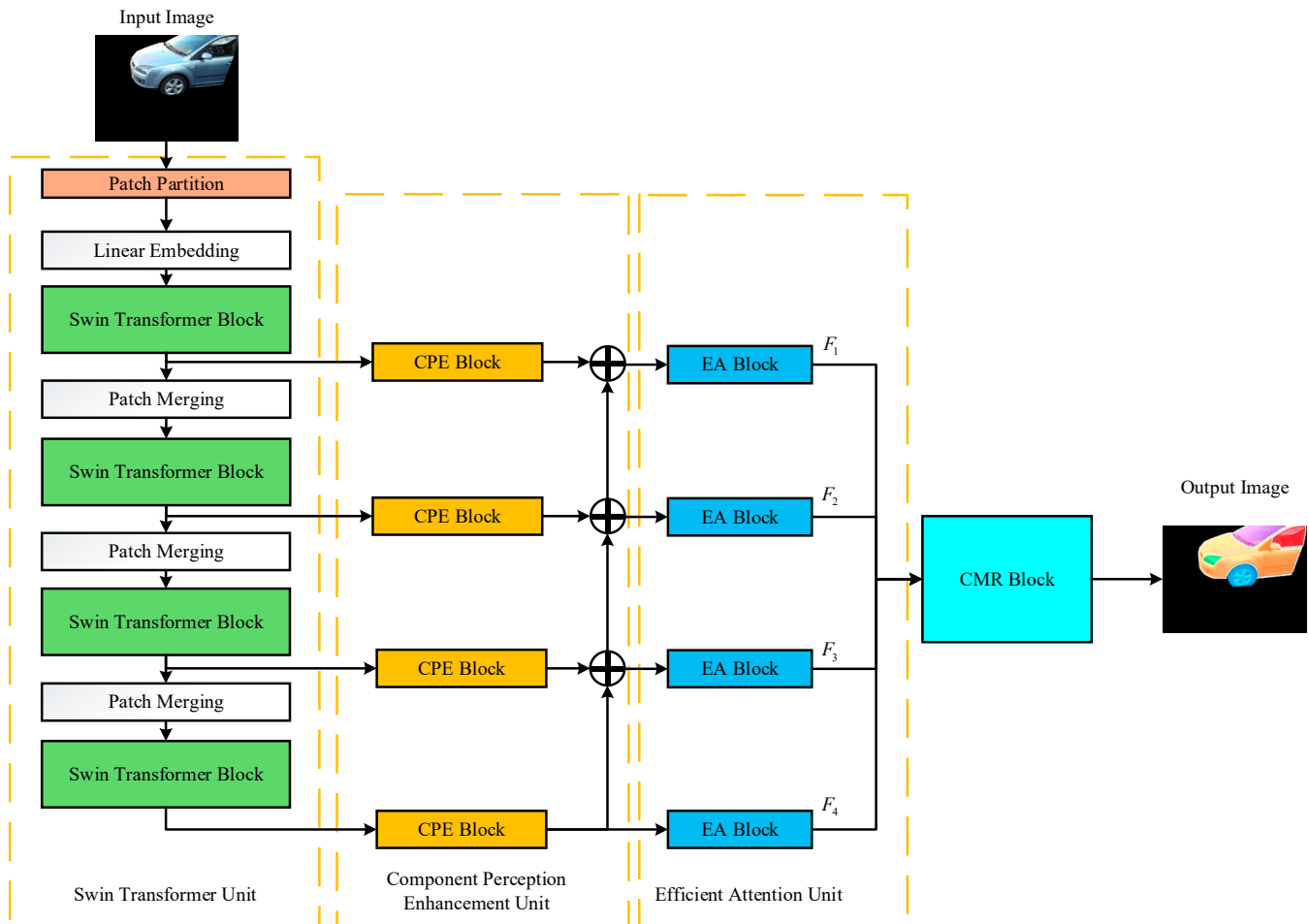


Figure 3. Vehicle component feature mask module structure.

As shown in Figure 3, the overall architecture of the vehicle component feature extraction and classification module includes four parts. First, the backbone net part uses the Swin Transformer unit to extract the vehicle component features of the saliency segmented image; then, the component perception enhancement (CPE) unit is used to extract the multi-scale features of the components, and the up-sampling fusion is performed through the deep to shallow component feature transfer to enhance the features of the weak contrast region. Then, it is sent to the efficient attention (EA) unit for filtering processing to reduce the interference of irrelevant regions, and the weighted regional features are fused to obtain the features of each vehicle component. Finally, a Mask Decoder is introduced to remove the damaged areas that cannot be classified as vehicle component, and the Component Mask Recognition (CMR) unit is used to determine whether these parts features belong to vehicle parts features, and finally, the vehicle is identified.

2.2.1. Swin Transformer Unit

The Swin Transformer unit is shown in Figure 3, which is composed of the Patch Partition layer, linear Embedding layer, Patch Merging layer, and Swin Transformer Block. The Patch Partition layer can divide the input vehicle saliency segmentation image into fixed-size image blocks and the linear embedding layer projects these segmented images onto the specified dimensions to form sub-images so that a complete image is encoded into a sequence. Then, it is input into the Swin Transformer Block to extract the regional features of the target vehicle parts, and the regional features of the target vehicle image can be obtained. In the second step, the Patch Merging layer is used to fuse the feature images of vehicle parts.

Two self-attention mechanisms, W-MSA and SW-MSA, are used in the Swin Transformer block, as shown in Figure 4.

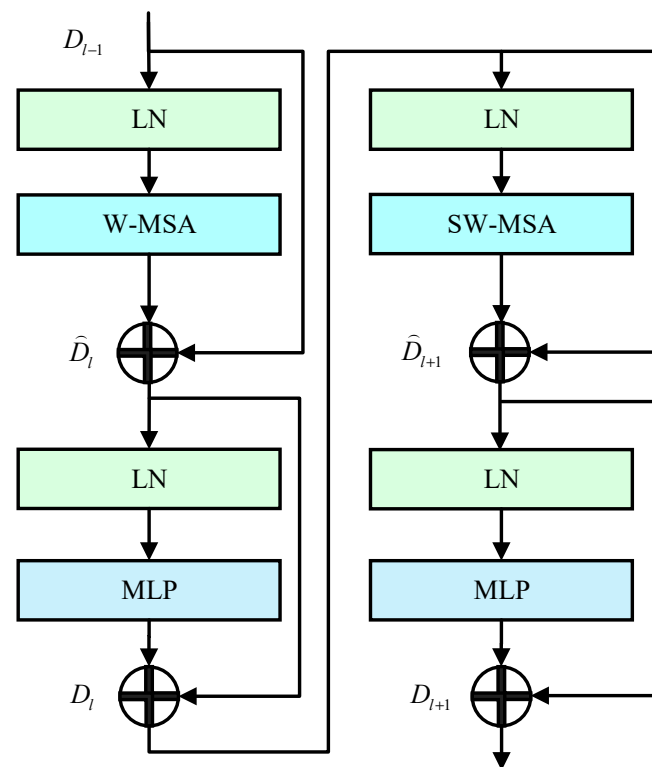


Figure 4. Swin Transformer block structure.

W-MSA is a window-based multi-head self-attention mechanism. In the feature extraction of vehicle components, its self-attention objects are the windshield, side glass, black wheels, blue license plates, and front and rear lights of the vehicle. For the extracted salient region, W-MSA divides the entire input salient region into several windows and performs matrix multiplication and SoftMax operations in each window, which can constrain the calculation range of the attention mechanism and reduce the calculation amount. The computational complexity is as follows:

$$O(W - MSA) = 4hwc^2 + 2M^2hwc \quad (3)$$

In Equation (3), h and w represent, respectively, the number of the patches in the height and width of images, M^2 is the number of the patches in the window, and C is the number of the channels in the window.

Since the number of patches inside the window is much smaller than the number of image patches, and the number of windows remains unchanged, the computational

complexity of the W-MSA is linearly related to the image size, which greatly reduces the computational complexity of the module.

W-MSA reduces the computational complexity while maintaining local dependence. W-MSA divides the feature map into several small Windows and sends them to MSA. The MSA calculation process is shown in the following steps:

- (1) The input features are mapped linearly to obtain query matrix Q , keyword matrix K , and value matrix V .
- (2) The score matrix S between vectors of different input elements is calculated according to Equation (4). The score determines the degree of attention of the encoding at the current position to other vectors.

$$S = Q \cdot K^T \quad (4)$$

- (3) Each element is normalized according to Equation (5) to enhance the stability of the gradient.

$$S' = S / \sqrt{d_k} \quad (5)$$

- (4) Perform a Softmax operation on S' to get each element of the attention matrix obtained by the operation of Equation (6).

$$a_{i,j} = \frac{e^{S'_{i,j}}}{\sum_{u=1}^N e^{S'_{i,u}}}, 1 \leq i \leq N, 1 \leq j \leq N \quad (6)$$

In Equation (6), $a_{i,j}$ represents the correlation between the i -th element and j -th element in input X . This step converts the scores between pixels into probabilities.

Based on the above expression, the MSA calculation sub-step can be summarized into a functional Equation (7), which is expressed as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d}} + B\right) \cdot V \quad (7)$$

In the Equation (7), $Attention$ denotes the self-attention matrix, d represents the vector dimension of the mapping matrix Q and K , and B is the bias matrix.

Although W-MSA can reduce computational complexity and calculate vehicle component features, there is a lack of information exchange between the non-coincident windows. Therefore, the next step is to enter the SW-MSA layer through a MLP with a GELU activation function. The role of the layer is to introduce a cross-window connection through a movable window so that the information between different local windows is integrated and the ability to extract global features is improved. In addition, a Layer Norm layer is introduced before each MSA layer and MLP layer, and residual joins are used overall.

The calculation process of the Swin Transformer Block is shown in Equation (8).

$$\begin{cases} \widehat{D}_l = W - MSA[LN(D_{l-1})] + D_{l-1} \\ D_l = MLP[LN(\widehat{D}_l)] + \widehat{D}_l \\ \widehat{D}_{l+1} = SW - MSA[LN(D_l)] + D_l \\ D_{l+1} = MLP[LN(\widehat{D}_{l+1})] + \widehat{D}_{l+1} \end{cases} \quad (8)$$

In the Equation (8), D_l and D_{l+1} are the regional features of the output vehicle components at different levels, $\widehat{D}_l$ and $\widehat{D}_{l+1}$ represent the output of (S)W-MSA and l th MLP. MLP is a multi-layer perceptron, LN is normalization operation, $W - MSA$ is the window attention mechanism, and $SW - MSA$ is the moving window attention mechanism.

2.2.2. Component Perception Enhancement Unit

To deal with the problem that the weak contrast of the vehicle component area in the acquired image affects the feature extraction, the component perception enhancement unit is introduced to enhance the vehicle component features, as shown in Figure 5:

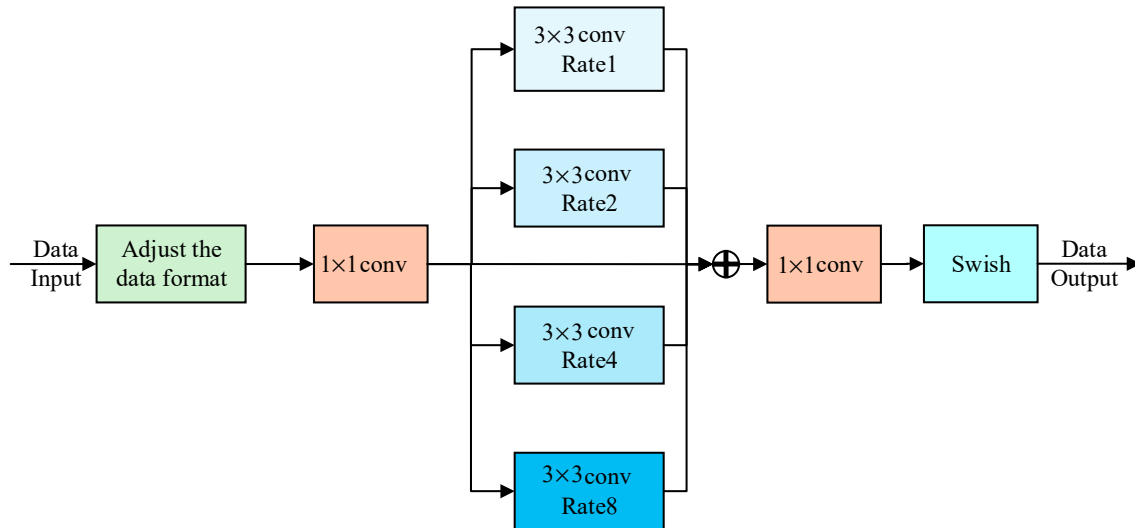


Figure 5. Component perception enhancement unit.

In Figure 5, the receptive field expansion is carried out by introducing dilated convolution to improve the response performance of the component perception enhancement unit to the weak contrast of vehicle components. Due to the particularity of the output feature data of the Swim Transformer unit, the data format needs to be adjusted first. Then, 1×1 convolution is used to reduce the dimension of the feature image to reduce the amount of parameter calculation. Then put the features into the dilated convolution with different magnifications. The magnifications are 1, 2, 4, and 8, and the size of the convolution kernel is 3×3 . The features on different branches are fused to generate the enhanced features, and 1×1 convolution is used to increase the dimension to obtain the feature map. Finally, the feature map is activated by the Swish function as the output feature data.

The introduction of dilated convolution in this module can make the net pay more attention to the component areas with weak vehicle contrast. The formula of dilated convolution is as follows:

$$f_o(m, n) = \sum_{i=1}^m \sum_{j=1}^n f_i(m + r \cdot i, n + r \cdot j) \cdot \omega(i, j) \quad (9)$$

In Equation (9), $f_o(m, n)$ is the element of the output feature map, $f_i(m + r \cdot i, n + r \cdot j)$ is the element of the input feature map, $\omega(i, j)$ is the element of the convolution kernel, m and n are the height and width of the convolution kernel, and r is the expansion rate.

2.2.3. Efficient Attention Unit

In the process of feature extraction of component regions, there may be a problem of component over-segmentation, which will affect the classification effect of target vehicle components, resulting in misclassification and missed classification. In order to reduce the interference of over-segmentation of vehicle components and enhance the component classification ability of the module, an efficient attention (EA) unit is designed, as shown in Figure 6.

The efficient attention mechanism uses 1D horizontal global average pooling (X Avg Pool) and 1D vertical global pooling (Y Avg Pool) to extract feature information. Then, the features are concatenated into a feature map using *Concat*, and 1×1 convolution is

used to obtain the dependency. After that, the split operation is performed along the X and Y directions, and the number of channels is restored by convolution. The Sigmoid nonlinear activation is used to realize the re-weighting of the regional features of the vehicle components and enhance the module's attention to the real components of the vehicle.

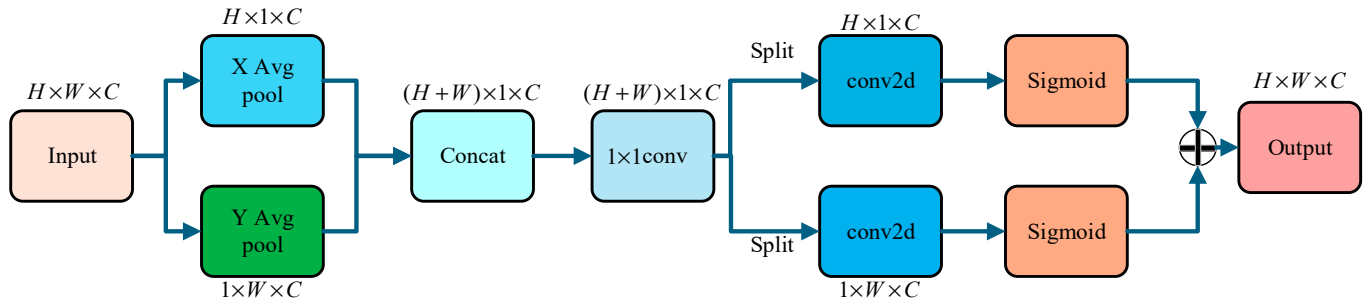


Figure 6. Efficient attention unit.

2.2.4. Component Mask Recognition Unit

In the face of the problem of foreground occlusion, such as mutual occlusion between vehicle targets and occlusion caused by non-target objects, it will affect the recognition of target vehicles. Therefore, to improve the recognition accuracy of the target vehicle under partial occlusion conditions, according to the extracted vehicle component features, a Mask Decoder is introduced to remove the damaged areas that cannot be classified as vehicle components. The component mask recognition (CMR) unit is designed to determine whether these component features belong to the vehicle component features, and finally, the vehicle is identified. The overall structure of the unit is shown in Figure 7.

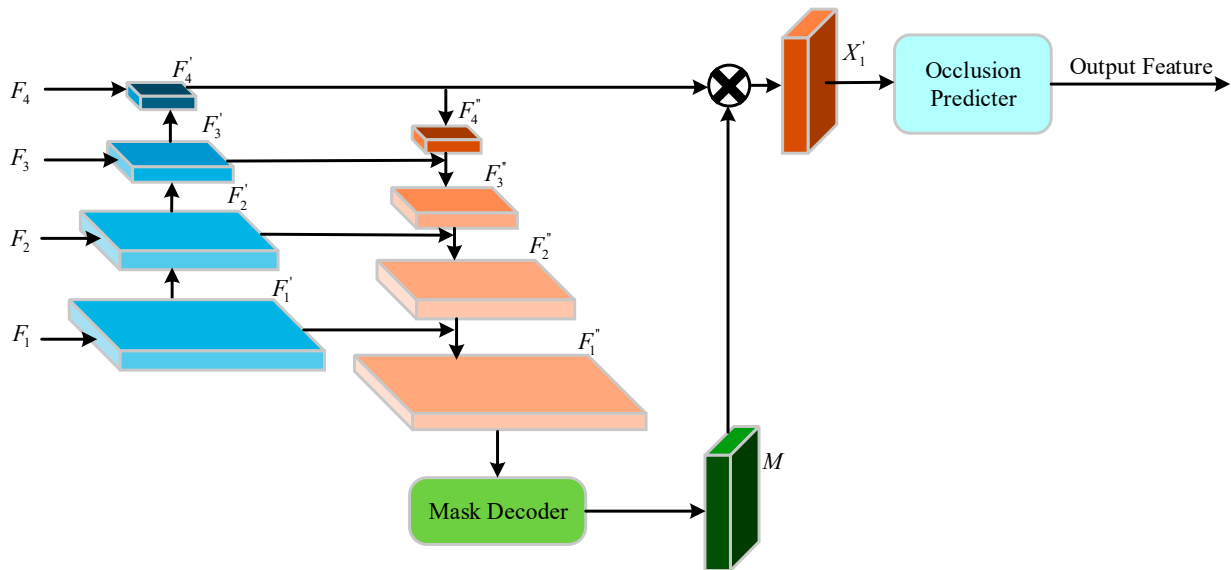


Figure 7. Component mask recognition unit.

In Figure 7, the CMR unit is composed of a multi-scale feature extraction block and a recognition block. Firstly, it uses the mid-deep features F_1' , F_2' to capture local and global information from the vehicle component feature input module obtained above. Then, it is sent to the Mask Decoder for accurate learning of feature masks. The expression equations of F_1' and F_2' are as follows:

$$F_1'' = \text{conv}(\text{upsample}(\text{conv}(F_1')) + \text{upsample}(\text{conv}(F_2')))) \quad (10)$$

$$F_2'' = \text{conv}(\text{upsample}(\text{conv}(F_2')) + \text{upsample}(\text{conv}(F_3')))) \quad (11)$$

The Mask Decoder structure is like an attention mechanism module. It allows the net to learn an appropriate weight matrix through a convolution operation to remove the features of occlusion damage in the feature map. The structure of the Mask Decoder is shown in Figure 8.

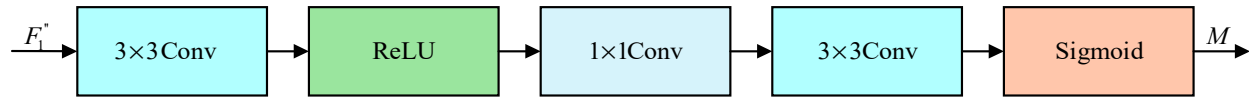


Figure 8. Mask Decoder structure.

The Mask Decoder structure decodes the occlusion information in the obtained feature map into a feature mask. It is expected to shield the occlusion damage feature elements of F_4' by an element-by-element product to generate feature X_1' for subsequent recognition.

3. Experiment and Result Analysis

3.1. Experimental Data Set

In this paper, the PASCAL VOC2012 data set, Vehicle Occlusion data set, UA-DETRAC data set, and self-collected real vehicle data set are used to verify the vehicle detection effect of the proposed algorithm. PASCAL VOC2012 is a standard data set that contains a variety of vehicles. The Vehicle Occlusion data set is a synthetic occlusion data set, in which the target object randomly overlaps with two, three, or four unrelated occlusions. UA-DETRAC is a challenging benchmark for real-world multi-target detection and multi-target tracking, it consists of 10 h of video captured by a Cannon EOS 550D camera at 24 different locations in Beijing and Tianjin, China. This paper only trains the vehicle targets in the data set.

During the experiment, 6713 car images of the three data sets were used as training sets, and the number of training sets was expanded by image enhancement processing methods such as scaling, rotation, and mirroring to further improve the performance of the module. The test uses 500 images collected by ourselves, and the vehicle target in the image has random occlusion. And 748 vehicle images in three public data sets form a data set for module verification.

Based on the above training data set and test data set, the accuracy rate, mIoU, standard average precision (AP, average precision), and mean average precision (mAP, mean average precision) were used as detection and evaluation indicators to evaluate the performance of the VSRS-VC FEC recognition net designed in this paper.

The vehicle components are divided into six categories, as shown in Table 1.

Table 1. Vehicle components category.

No.	Component Name	Component Label Color
1	Front window	purple
2	Side window glass	red
3	Rear window	brown
4	Wheel	blue
5	Lamp	green
6	License plate	dark blue

3.2. Experimental Environment and Parameter Configuration

The computer used in the experiment is configured with a Windows 11 operating system, NVIDIA GeForce RTX 4060 Laptop GPU, i5-11400H CPU, 16GB memory, Python3.10 computer language, and a Tensorflow2.6 deep learning framework to run the VSRS-VC FEC vehicle recognition net designed in this paper.

In the training phase, the batch size is set to 4, and the stochastic gradient descent algorithm is used to update the recognition net weight. The initial learning rate is 0.008, and

the learning rate is reduced to 10% for every 16 epochs. The maximum iterative training is set to 24 epochs.

3.3. Ablation Experiment

To verify the effectiveness of the vehicle recognition method in this paper, the ablation experiments were carried out on the two units of the component perception enhancement unit and the efficient attention unit in the vehicle component feature extraction and classification module designed in Section 2.2 and the experimental verification was carried out on the PASCAL VOC2012 data set. Table 2 shows the changes in vehicle recognition performance after the addition of two units, among which “×” indicates that the Vehicle Component Feature Mask module does not introduce this unit, and “√” indicates that the Vehicle Component Feature Mask module introduces this unit.

Table 2. Ablation experimental results.

Component Perception Enhancement Unit	Efficient Attention Unit	PASCAL VOC2012		
		Precision/%	Accuracy/%	mAP/%
×	×	87.9	82.8	88.8
√	×	89.5	81.8	91.1
×	√	88.1	86.0	92.5
√	√	90.8	88.7	92.9

The real-time analysis of these modules shows that the CPE unit has the greatest impact on the success rate, which makes the extraction accuracy rate increase by 1.6% compared with the original feature extraction net, and the EA unit has the greatest impact on the accuracy rate compared with the original feature extraction net of 3.2%. When the RFE unit and EA unit are applied to the module at the same time, the module obtains a higher mAP value, reaching 92.9%. It shows that the two improvements can improve the module performance at the same time. Through ablation experiments, it is proved that the extraction algorithm in this paper has better performance. The mAP reaches 92.9%, which is 4.1% higher than the original module, and the accuracy is improved compared with the original module. In summary, the recognition method in this paper has improved the performance of feature extraction.

3.4. Comparative Experiments

To verify the advanced characteristics of the VSRS module backbone network designed in Section 2.1 of this paper, the backbone network of this module is compared with the currently realized and representative mainstream instance segmentation models, including DetNet, which does not use pre-trained classifiers, and vgg16 and resNet101, which use pre-trained classifiers as backbone networks, respectively. The parameters and environment configured during the experiment were consistent. The performance pairs are shown in Table 3.

Table 3. The experimental results of VSRS module backbone performance comparison.

Backbone	mAP/%	mIoU/%	Accuracy/%	Precision/%	Recall/%
Vgg16	79.57	82.19	80.4	81.09	80.01
ResNet101	81.44	85.69	83.74	82.71	82.67
DetNet	83.29	84.58	84.25	84.91	82.91
The method proposed in this paper	84.4	86.81	86.12	86.69	84.16

As can be seen from Table 3, mAP, mIoU, accuracy, precision, and Recall of the backbone network using the proposed method are 4.83%, 4.62%, 5.72%, 5.6% and 4.15% higher than that of vgg16, respectively. They are 2.96%, 1.12%, 2.38%, 3.98% and 1.49%

higher than ResNet101. It is 1.13%, 2.23%, 1.87%, 1.78% and 1.25% higher than DetNet, respectively. It can be concluded that the performance of vgg16 is the lowest of all the above algorithms. ResNet101 algorithm and DetNet algorithm have better performance. The method module in this paper is superior to the above four algorithms in recognition accuracy and segmentation accuracy, and shows better performance.

In order to verify the advanced nature of the vehicle component feature extraction classification module designed in Section 2.2, the module is compared with the currently implemented and representative mainstream instance segmentation modules, including SOLOv2 [25], Mask R-CNN [26], Cascade Mask R-CNN [27], QueryInst [28]. The parameters and environment configured during the experiment are consistent. The performance comparison is shown in Table 4.

Table 4. The experimental results of VCFM module performance comparison.

Algorithm	mAP/%	mIoU/%	Accuracy/%	Precision/%	Recall/%
SOLOv2	84.2	79.98	83.40	83.95	84.01
Mask R-CNN	83.29	82.40	86.87	85.94	85.76
QueryInst	89.4	84.58	88.65	88.29	87.95
Cascade Mask R-CNN	86.71	83.52	88.17	88.04	87.62
The method proposed by this paper	88.72	85.59	89.82	89.47	89.16

It can be seen from Table 4 that the mAP, mIoU, accuracy, Precision, and Recall of the proposed method are 4.52%, 5.61%, 6.42%, 5.52%, and 5.15% higher than SOLOv2, respectively, and 5.43%, 3.19%, 2.95%, 3.53%, and 3.4% higher than the CNN-based vehicle segmentation algorithm Mask R-CNN method, respectively. In addition, compared with the detection accuracy and segmentation accuracy of other advanced modules, it can be concluded that the overall detection accuracy and segmentation accuracy of SOLOv2 are the lowest among all the above algorithms. The Cascade Mask R-CNN algorithm and QueryInst algorithm have relatively good detection accuracy and segmentation accuracy. The module proposed exceeds the above four algorithms in terms of recognition accuracy and segmentation accuracy, showing better performance. Therefore, the proposed method can better extract and utilize the features of the target, and the module segmentation performance is better.

3.5. Comparison Experiment of Occlusion Vehicle Recognition Net Results

To evaluate the actual recognition effect of VSRS-VCFM and the currently realized representative mainstream recognition algorithms, a comparative experiment was designed, and the vehicle image was divided into four types according to its occlusion forms: no occlusion, field of view occlusion, non-target occlusion, and inter-target occlusion, and prove the advantages of VSRS-VCFM. The experimental results are shown in Table 5.

Table 5. Experimental results of performance comparison of VSRS-VCFM in recognizing traffic vehicle scenes on UA-DETRAC.

Algorithm	mAP/%	mIoU/%	Accuracy/%	Precision/%	Recall/%
Mask R-CNN	70.41	70.54	73.06	72.34	70.14
QueryInst	70.99	71.78	73.56	72.12	73.99
Cascade Mask R-CNN	71.08	71.02	72.91	73.47	73.1
The method proposed in this paper	74.27	72.7	76.37	74.92	75.05

It can be seen from Table 5 that the Mask R-CNN on traffic vehicle scenes performance is the worst. The mAP, mIoU, accuracy, Precision, and Recall of the proposed method are 3.28%, 0.92%, 2.81%, 2.58%, and 1.06% higher than QueryInst, respectively, and 3.19%, 1.68%, 3.46%, 1.45%, and 1.95% higher than the Cascade Mask R-CNN method, respectively.

The superiority of this method is proved. In summary, the module proposed in this paper exceeds the above three algorithms in recognition accuracy and segmentation accuracy, showing better performance. Therefore, the method can better identify the vehicle parts, so as to realize the identification of shielding vehicles.

A set of representative vehicle target images in different states are selected. The algorithm is compared with the representative mainstream occlusion target recognition algorithms Cascade R-CNN algorithm and QueryInst algorithm. The results are shown in Figure 9. Figure 9a shows the actual images of occluded vehicles. The detection and segmentation results of the Cascade R-CNN are shown in Figure 9b, the detection and segmentation results of the QueryInst model are shown in Figure 9c, and the detection and segmentation results of the proposed algorithm are shown in Figure 9d.

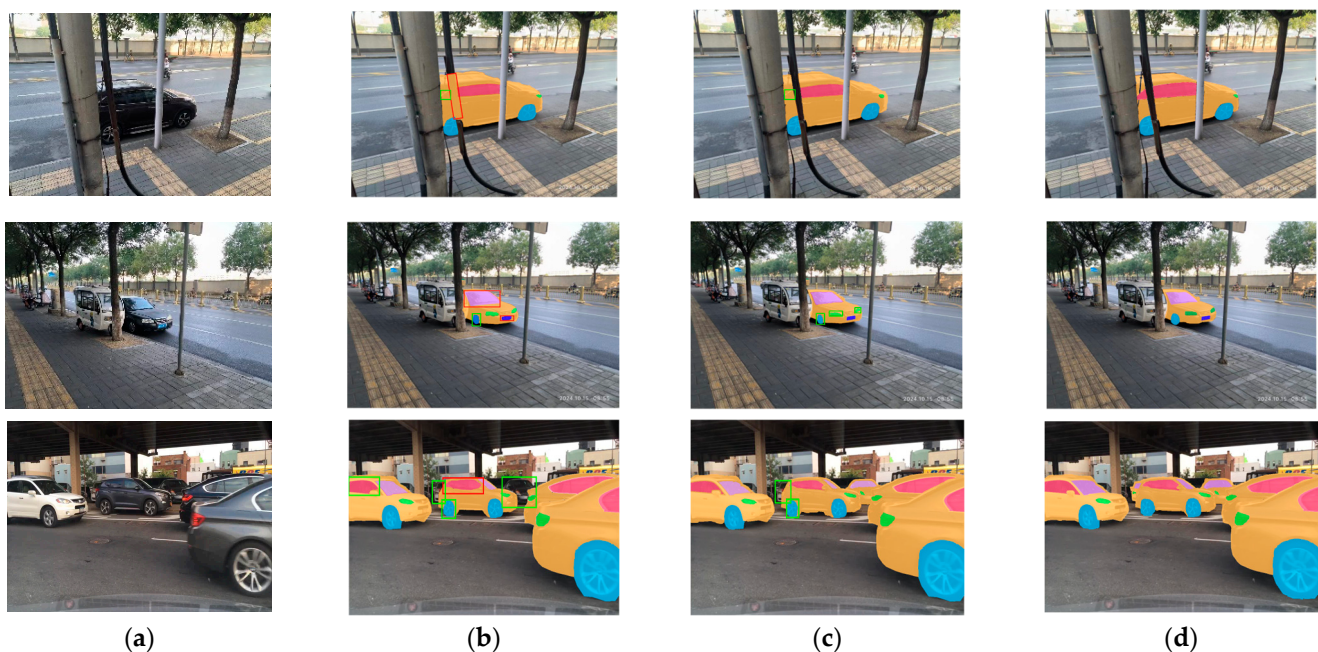


Figure 9. Visual comparison results. (a) Vehicle image 1, (b) Cascade R-CNN segmentation result based on (a), (c) QueryInst segmentation result based on (a), (d) segmentation result of the method proposed in this paper based on (a).

The test results are shown in Figure 9. The missed part is marked with a green box and the over-segmentation part is marked in red boxes. In the first set of images, the Cascade R-CNN model failed to detect some components and incorrectly classified some obstacles as components, resulting in missed detection and false detection. The QueryInst model does not detect some components, and the effect of vehicle contour detection is poor. Due to the introduction of the VSRS module and VCFM module, the model can segment the vehicle more accurately and can correctly identify the missing areas in the previous two models. In the second set of images, the Cascade R-CNN model has a poor detection effect on vehicle components. QueryInst also has a poor detection effect on the vehicle contour. In the model testing results of this paper, the accurate segmentation of vehicle parts and vehicles can be realized. In the third group of images, the Cascade R-CNN model appears to recognize vehicle leakage, and the detection effect of the vehicle contour is poor. QueryInst also has some cases of missing recognition. The results of model inspection in this paper can better realize automotive parts and segmentation.

4. Conclusions

In the process of vehicle detection, due to the complex background, illumination changes, and occlusion, the accuracy of vehicle target segmentation from the background is poor, and the recognition accuracy is poor. This paper proposes a method of occluded

vehicle target detection based on vehicle component features, uses the U-net module, and introduces the residual U-shaped block so that the module can find the vehicle target area from the complex image. It introduces a vehicle component feature extraction module based on the Swin Transformer to effectively solve the problem of misdetection and missed detection of occluded vehicle targets caused by insufficient feature extraction. It combines the Swin Transformer module, receptive field expansion module, and efficient attention module and efficiently obtains the vehicle target component features and realizes the recognition of vehicle targets after the extracted feature labels and images are sent to the classifier trained by CMR. The experimental results show that the proposed method can effectively solve the problem of misdetection and missed detection of occluded vehicle targets due to insufficient feature extraction, and the recognition effect of occluded vehicles is better than other advanced methods.

In practice, vehicle data are easily subject to different interference, so more features need to be considered to enhance model recognition, such as color features. In the next step, the color feature recognition part will be added on the premise of ensuring the speed of the model detection so as to improve the detection accuracy of the model and make it more suitable for practical application.

Author Contributions: Conceptualization and methodology, software and validation, writing—original draft, H.H.; conceptualization and methodology, writing—review and editing, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Science and Technology Program of Shaanxi Province, grant number 2023-YBGY-342, and the National Natural Science Foundation of China, grant number 62073256.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Hu, L.W.; Lei, G.Q.; Zhao, X.T. Study on the traffic accident risk propagation and control in urban congestion environment. *J. Saf. Environ.* **2023**, *23*, 2809–2818.
2. Xie, K.Q. Vehicle identification system based on MATLAB. *Electron. Technol. Softw. Eng.* **2017**, *9*, 65.
3. Reddy, N.D.; Vo, M.; Narasimhan, S.G. Occlusion-Net: 2D/3D occluded key point localization using graph networks. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7326–7335.
4. Chen, Y.J.; Xu, F.; Chen, G.D. Point cloud 3D object detection method based on density information-local feature fusion. *Multimed. Tools Appl.* **2024**, *83*, 2407–2425. [[CrossRef](#)]
5. Sun, F.W.; Li, C.Y.; Xie, Y.Q. Review of Deep Learning Applied to Occluded Object Detection. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 1243–1259.
6. Wang, S.N.; Mei, Y.U.; Jiang, G.Y. Review on vehicle detection and tracking techniques based on video processing in intelligent transportation systems. *Appl. Res. Comput.* **2005**, *22*, 9–14.
7. Tian, Y.; Luo, P.; Wang, X. Deep learning strong parts for pedestrian detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1904–1912.
8. Zhang, S.; Wen, L.; Bian, X. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Munich, Germany, 2018; pp. 637–653.
9. Yu, X.H.; Mao, S.C.; Wang, J.B. A two-stage occlusion detection method based on improved MAE and YOLOv5. *Inf. Technol. Informatiz.* **2023**, *8*, 31–38.
10. Devries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
11. Hinton, G.E.; Srivastava, N.; Krizhevsky, A. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
12. Yun, S.; Han, D.; Oh, S.J. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6022–6031.

13. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 1–14. [[CrossRef](#)]
14. Wang, Y.; Tao, X.; Qi, X. Image inpainting via generative multi-column convolutional neural networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems, Montreal, Canada, 3–8 December 2018; pp. 31–40.
15. Zhao, L.; Mo, Q.; Lin, S. UCTGAN: Diverse Image Inpainting Based on Unsupervised Cross-Space Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5741–5750.
16. Yu, J.; Lin, Z.; Yang, J. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514.
17. Zhou, C.L.; Yuan, J.S. Multi-label learning of part detectors for occluded pedestrian detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3486–3495.
18. Ren, S.; He, K.; Girshick, R. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015; pp. 91–99.
19. Liu, T.J.; Gu, N.J.; Zhang, X.C. Fine-grained recognition of vehicle model using multi-scale feature fusion CNN. *Comput. Eng. Appl.* **2018**, *54*, 154–160.
20. Bai, S.; Liu, Z.; Yao, C. Classify vehicles in traffic scene images with deformable part-based models. *Mach. Vis. Appl.* **2018**, *29*, 393–403. [[CrossRef](#)]
21. Li, G.J.; Hu, J.; Ai, J.Y. Vehicle Detection Based on Improved SSD Algorithm. *Comput. Eng.* **2022**, *48*, 266–274.
22. Deng, Z.; Sun, H.; Zhou, S. Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3652–3664. [[CrossRef](#)]
23. Ronneberge, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Munich, Germany, 2015; pp. 234–241.
24. Liu, Z.; Lin, Y.; Cao, Y. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
25. Wang, X.L.; Zhang, R.F.; Kong, T. SOLOv2: Dynamic and fast instance segmentation. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; p. 1487.
26. Luo, Y.; Zhao, H.; Xu, F.L. The Application of Improved Mask R-CNN in Vehicle Instance Segmentation. *J. Qingdao Univ.* **2023**, *38*, 27–34.
27. Cai, Z.W.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1483–1498. [[CrossRef](#)] [[PubMed](#)]
28. Fang, Y.X.; Yang, S.S.; Wang, X.G. Instances as queries. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual Conference, 11–17 October 2021; pp. 6890–6899.

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.