electronics

Article

Vehicle Logo Detection Method Based on Improved YOLOv4

Xiaoli Jiang [1], Kai Sun [1,2], Liqun Ma [1], Zhijian Qu [1,*] and Chongguang Ren [1]

[1] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[2] Zibo Special Equipment Inspection Institute, Zibo 255000, China
* Correspondence: zhijianqu@sdut.edu.cn

Abstract: A vehicle logo occupies a small proportion of a car and has different shapes. These characteristics bring difficulties to machine-vision-based vehicle logo detection. To improve the accuracy of vehicle logo detection in complex backgrounds, an improved YOLOv4 model was presented. Firstly, the CSPDenseNet was introduced to improve the backbone feature extraction network, and a shallow output layer was added to replenish the shallow information of small target. Then, the deformable convolution residual block was employed to reconstruct the neck structure to capture the various and irregular shape features. Finally, a new detection head based on a convolutional transformer block was proposed to reduce the influence of complex backgrounds on vehicle logo detection. Experimental results showed that the average accuracy of all categories in the VLD-45 dataset was 62.94%, which was 5.72% higher than the original model. It indicated that the improved model could perform well in vehicle logo detection.

Keywords: vehicle logo; small object detection; DenseNet; deformable convolution; Vision Transformer

Citation: Jiang, X.; Sun, K.; Ma, L.; Qu, Z.; Ren, C. Vehicle Logo Detection Method Based on Improved YOLOv4. *Electronics* **2022**, *11*, 3400. https://doi.org/10.3390/electronics11203400

Academic Editors: Manohar Das and Byung Cheol Song

Received: 30 August 2022
Accepted: 19 October 2022
Published: 20 October 2022

1. Introduction

Vehicle information detection is an important branch of intelligent transportation systems. As one of the important pieces of information about vehicles, the vehicle logo has characteristics obvious and difficult to replace, so its recognition is of great significance. As another important information to identify the vehicle, license plate recognition is recognized as one of the most effective means of vehicle recognition and has been widely and successfully implemented [1]. However, the license plate is easily removed, occluded or tampered with. Therefore, only obtaining vehicle information through a license plate cannot meet the actual demand. Vehicle logos are key symbols of automobile manufacturers, carrying vital information about vehicle identification and are difficult to counterfeit. Consequently, an accurate identification of vehicle logos is useful for identifying fake-licensed cars, performing vehicle control queries, and tracking vehicle violations.

The existing vehicle logo recognition (VLR) methods are mainly divided into two categories: traditional handcrafted-feature-based methods and deep convolutional neural network (CNN)-based methods. Traditional machine learning algorithms design different manual features to train detectors that are easy to understand and implement. Histogram of oriented gradient (HOG) [2], invariant moments [3], and scale-invariant feature transform (SIFT) [4] are the most commonly used traditional features. However, a special detector can only be designed for a class of vehicle logo characteristics. It can detect the type of logo when relatively simple but is not enough in a variety of complex environments to identify so many kinds of vehicle logos. In recent years, object recognition based on deep learning has been dominant in computer vision tasks, and it has applied some deep learning methods to vehicle logo recognition. Huang et al. [5] migrated the convolutional neural network to the vehicle logo recognition task, and introduced the pretraining strategy to obtain breakthrough results in a large-scale 10-class database. Yu et al. [6] proposed a multilayer pyramid network for vehicle logo recognition and achieved good detection results. Refs. [7,8] proposed a large dataset for vehicle logo detection and continuously improved

the accuracy and speed of vehicle logo detection and recognition on the benchmark model of object detection.

Although the deep learning model has achieved good results in the field of vehicle logo recognition, there are still some problems in existing models' algorithms in vehicle logo recognition and classification. Firstly, the vehicle logo makes up only 0.1% to 1% of the overall image. The small size has a significant impact on the feature extraction of the target. Further, due to the effects of light intensity, shooting angle, vehicle location and a large amount of background interference information in the natural environment, the detector must meet higher standards than mainstream detectors. Finally, the irregular shape of the vehicle logo, especially the letters such as JEEP and HAVAL whose shape is more abnormal, brings a great difficulty in detection.

To solve the above problems, a vehicle logo recognition method based on the improved YOLOv4 [9] model is presented in this paper. This method is suitable for common vehicle logo detection under complex backgrounds, as it has a better generalization ability and robustness. In particular, an effective solution is proposed for the recognition of small-scale objects, irregular shapes, and complex backgrounds.

The main contributions of this work are summarized as follows:

- Aiming at the characteristics of small vehicle logo objects, the backbone feature extraction network of YOLOv4 is improved;
- The shape of the logo is complex and irregular, which leads to our model reconstructing the neck part of YOLOv4;
- A detection head named CT-Head is designed. CT-Head is suitable for reducing the influence of complex backgrounds on vehicle logo detection.

2. Related Work

In the early stage of research on vehicle logo recognition, the traditional handcrafted-features-based methods employed histogram, texture, moment invariants, and other traditional features to describe the vehicle logo. After extracting the features, the classification and recognition of the vehicle logo were achieved through machine learning methods. Pan et al. [10] proposed a fast and reliable vehicle logo detection method, which included three steps: vehicle region detection, small ROI segmentation, and logo detection. Firstly, the improved AdaBoost algorithm was used to extract the vehicle region from the input image, and the ROI was segmented from the detected vehicle. Then, a two-stage cascaded classification scheme was proposed to locate the vehicle logo from the focused small ROI region. Thubsaeng et al. [11] presented a new method for identifying vehicle logos from a vehicle's front and rear views. This work was a two-stage method that combined CNN and a pyramid histogram of oriented gradients (PHOG) features. CNN was used to identify candidate regions, and then the PHOG and an SVM were used to verify them. Compared with previous methods of VLR, this method showed good robustness. Peng et al. [12] proposed a new feature representation strategy—random sparse distribution (SRSD)—for low-resolution and low-quality images collected in intelligent transportation systems. It used the correlation between random sparse sampling pixel pairs as image features to describe the distribution of gray images statistically. At the same time, a creative classification algorithm, multiscale scanning, was proposed to reduce the influence of the propagation error in traditional methods by separating positioning and classification. Zhao et al. [13] proposed to increase the moment invariant feature to improve the recognition accuracy of vehicle logo recognition under weak light conditions. The vehicle logo was coarsely located based on the characteristics of the license plate and finely located with image processing technology. Then, HU invariant moments were used to extract vehicle logo features to establish a database. Finally, the SVM training process was optimized to obtain the optimal vehicle classification model. Yu et al. [14] constructed an updated vehicle logo recognition dataset (HFUT-VL) and proposed a new VLR method based on the HFUT-VL. This method used the overlapping enhanced patterns of oriented edge magnitudes (OE-POEM) features to represent vehicle logos and used the CRC classifier to identify vehicle logos. OE-POEM

was a descriptor derived from their improved oriented edge magnitudes (POEM), using the rich texture and edge information of vehicle logos.

These traditional handcrafted-features-based methods were simple and efficient, but they involved only fewer types of vehicle logos. Therefore, many types of vehicle logos could not be identified. In addition, artificial design features have low portability and insufficient robustness, so they are not sufficient to meet practical application conditions.

In recent years, with the continuous development of deep learning technology, more and more deep learning models have been proposed. Deep learning models have achieved automatic and accurate vehicle logo recognition and have achieved good recognition results. Vehicle logo recognition based on deep learning does not require artificial design features but instead learn feature representations from vehicle logos. Huang et al. proposed a CNN model for VLR. Xia et al. [15] proposed a method that combined CNN with multitask learning to identify and predict vehicle logos. In order to accelerate the convergence of the multitask model, they employed an adaptive weight-training strategy. They expanded the Xiamen University vehicle logo recognition dataset and achieved a high detection accuracy on the dataset. Li and Hu [16] also proposed a method with a pretraining strategy. They used the Hadoop framework for data processing and trained a CNN model for VLR. Yu et al. [17] proposed a tow-stage framework based on the cascaded deep convolution network for detecting vehicle logos. This method did not depend on the license plate detection but directly detected the vehicle logo. The region proposal network generated regions that may contain the vehicle logo. These regions were then divided into the background and different types of vehicle logos by convolutional capsule networks. Ke and Du [18] proposed three vehicle-logo data-enhancement strategies to solve the problem of small areas and small datasets. For the problem of the small sample size, the cross-sliding segmentation method was adopted. A small border method was proposed to expand the vehicle logo area in the image, and a Gaussian distribution based vehicle logo segmentation method was proposed to enrich the vehicle logo difference in the image position. These optimization methods could better reflect the characteristics of the vehicle logo than the traditional way, and they achieved excellent results in some object detection frameworks. Liu et al. [19] proposed a vehicle logo recognition method based on enhanced matching, a constrained region extraction, and a single-shot feature pyramid detector (SSFPD) network. The first step was the constrained region detection based on Faster-RCNN, which was used to extract the position information of the car's head and tail. Then, in the training process, the data were enhanced by copying and pasting vehicle logos. Finally, the ResNext network was improved to obtain the SSFPD network to extract features and generate feature maps. They improved the method according to the characteristics of small targets and improved the accuracy of vehicle logo detection. However, the dataset they used only had 13 types of vehicle logos, with limited scenarios. Zhang and Yang proposed a lightweight network structure based on separable convolution based on the YOLO structure [20–22]. This model improved the real-time performance of vehicle logo detection. Yang et al. [7] proposed a new dataset (VLD-45) for vehicle logo detection. The new dataset presented several research challenges, involving small objects, shape deformation, low contrast, and so on. They also evaluated VLD-45 using existing classifiers and detectors and the results showed that the proposed dataset had a very important research value for small-object detection tasks. Subsequently, they proposed a new approach called multiscale vehicle logo detector (SVLD) [8], which was based on SSD [23]. The biggest feature of SSD was its ability to detect multiscale objects. Based on this feature, by changing the network structure, setting preset boxes, and performing multiple rounds of pretraining, the SVLD algorithm was designed for vehicle logo detection. SVLD achieved good effects in both detection accuracy and detection speed. However, for images with complex background, there were still false detections, which mistakenly believed that the background was the vehicle logo. Lu et al. [24] proposed a novel category-consistent deep learning framework for the accurate identification of vehicle logos. The framework consisted mainly of two parts. One part was a vehicle logo feature extraction CNN (called VLF-net) to extract hierarchical features automatically. The other part

was a category-consistent mask-learning (CCML) module. CCML could learn the category-consistent regions without knowing the exact logo area to help the network focus on the region. One prominent contribution of the framework proposed by them was that they no longer relied on the labeling of artificial boundary frames. However, the disadvantage was that only the logo could be recognized from the front image of the vehicle, which required a high quality for the input image. The above operations performed well on several public vehicle logo datasets.

Although deep learning methods have achieved good results in vehicle logo detection, the accuracy of small object recognition in complex environments remains to be improved. In recent years, more and more research has tended to explore the method of small object detection [25,26]. Research on vehicle logo detection methods is helpful to promote the development of small object detection.

3. Methodology

3.1. Model Overview

A vehicle logo recognition model based on YOLOv4 is presented in this paper. YOLOv4 is a simple and efficient one-stage object detection model, the fourth version of the YOLO series. This model mainly includes three modules: backbone (Darknet53), neck (path aggregation network PAN and feature pyramid FPN) and head (YOLOHead). The structure of the YOLOv4 model is shown in Figure 1. Our model adapts and improves the three modules of YOLOv4 according to the characteristics of vehicle logos in natural environments. The structure of the enhanced model can be seen in Figure 2.

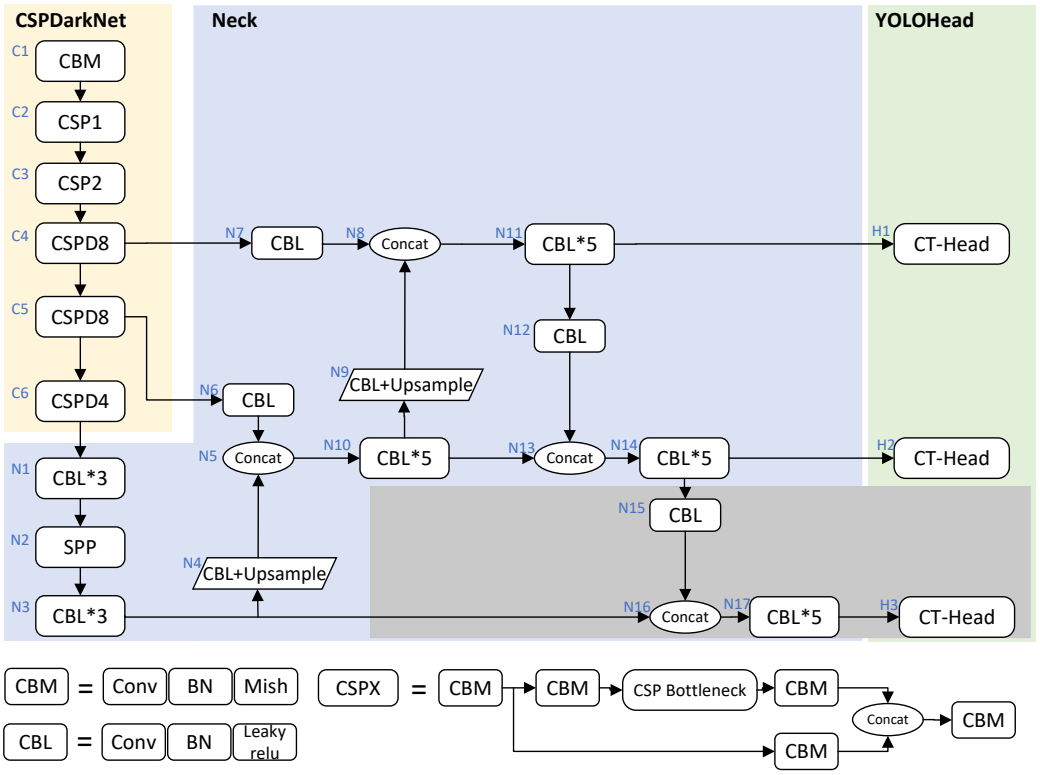


Figure 1. Structure of the YOLOv4 model.

In Figures 1 and 2, the number of each block is marked with blue numbers on the left corner side of the block. The improved part is marked with pink in Figure 2. Furthermore, the structure of the gray area in Figure 1 is deleted. As shown in Figure 2, in general, the improvements of the model include the following: a new output layer is added after C3 and it retains only two YOLOHeads; the residual dense connection structure (CSPD8, CSPD4) is employed in C4–C6 to replace the original CSP structure in the backbone; the

convolution blocks in feature pyramids are replaced by residual deformable convolution structures (CRD) and a new detection head structure, CT-head, is proposed. Specifically, what motivates us to make these improvements to the model and how to improve the model are described below.

Firstly, the proportion of the vehicle logo on the entire map is small, so vehicle logo recognition belongs to the small object detection task. According to this characteristic, on the basis of the original three output layers of YOLOv4, a shallow output layer (N14–N19 in Figure 2) was added. This method can supplement the shallow information about the feature pyramid fusion process and reduce the loss of vehicle logo information. Moreover, the images taken in real life contain vehicles at various angles, and the location of the vehicle logo is different.

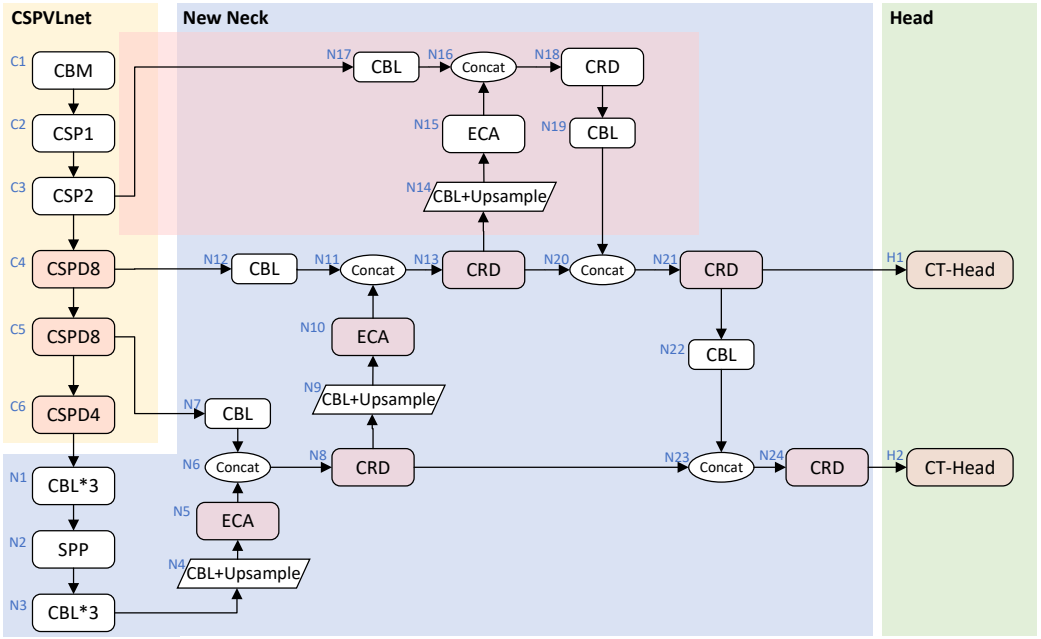


Figure 2. Structure of the improved YOLOv4 model.

Secondly, the vehicle logo shape is complex and changeable, especially since some alphabetic logo shapes are highly complex. Most vehicle logos are located on the front grille of the vehicle, and the color of the logo is similar to that of the grille. It is difficult to learn the vehicle logo's edge characteristics under complex and similar background information by using the sampling points obtained from the conventional convolution rule. The deformable convolution kernel can better extract input features of vehicle logos by adjusting its shape during sampling. Therefore, deformable convolutions were used to replace some of the ordinary convolutions to reconstruct the neck of YOLOv4, as shown in the blue area in Figure 2.

Finally, in recent years, with the increasing number of vehicles, the position of the logo of vehicles has progressively not been limited to the head part. In particular, the images to be detected are generally images of road vehicles, images of monitoring vehicles, and billboards of vehicles. The issue is not only the interference of the vehicle parts on the detection of the vehicle logo, but also the interference of the environment around the vehicle on the extraction of the vehicle logo information. Aiming at the problem of complex background information, the efficient channel attention for deep convolutional neural networks (ECA-Net) [27] was added to the N5, N10, and N15 parts of Figure 2 to pay more attention to the vehicle logo features. At the same time, to prevent the network from focusing only on local features, a convolutional transformer block was introduced to improve the head part. A convolutional transformer block is good at both local feature extraction and global receptive field.

3.2. The CSPVLnet

The proposed CSPVLnet retained the original CSPResNet design in sections C2 and C3 of CSPdarknet53. The original structure of the C4–C6 sections was replaced with CSPDenseNet. The number of dense connection layers in the new structure was equal to the number of residual structures in BottleNet in the original structure. The CSPDenseNet [28] structure is shown in Figure 3, taking CSPD4 as an example and changing the number of nodes in the Dense Block to 8 is the CSPD8 structure diagram.

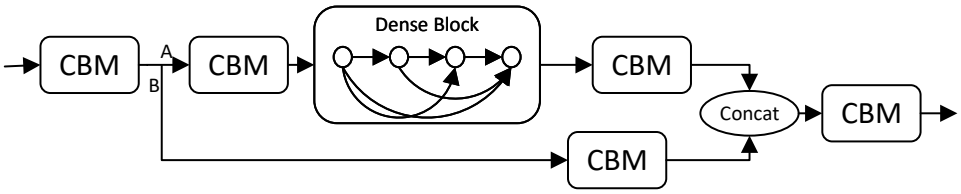


Figure 3. The CSPDenseNet structure (CSPD4).

As shown in Figure 3, the CSP structure divided the input characteristic data into two parts by convolution. The trunk part A adjusted the number of channels through a 1 * 1 convolution. After several dense connection blocks, a 1 * 1 convolution was used to integrate the channel characteristics. Then, the residual edge part B and the trunk part A were stacked in the channel dimension, and the channel information was fused through 1 * 1 convolution.

Different from the CSPResNet structure, CSPDenseNet allows each layer to receive the mappings of previous functions. CSPDenseNet makes the network more compact, and features and gradients transfer more efficiently. This structure extract features easily, and the network is easy to train. Some CSPResNet structures in CSPDarkNet were replaced by CSPDenseNet, and the proposed CSPVLnet combining the residual structure and dense connection structure was obtained. This network structure made the forward transmission of data information in the feature extraction network more diversified and the transmission capacity stronger. Therefore, CSPVLnet improved the reuse rate of small object features, which helped to improve the network learning ability and the accuracy of vehicle logo detection.

3.3. New Neck Structure

3.3.1. Additional Layer for Small Objects

Dealing with the problem of feature scale for detecting the vehicle logo is very important. The lack of superficial information leads to the missed detection of small objects, and deep semantic features may reinforce the characteristics of shallow spaces. The combination of shallow features and deep features can better extract the characteristics of small objects. Therefore, a shallow output layer was added based on the original three output layers of YOLOv4. It was integrated into the original feature pyramid to supplement the shallow position information in the fusion result of the feature pyramid. The shallow location information was added, and deep semantic features were fused by this method. In addition, most vehicle logos are small objects. In order to prevent too many parameters, the 13 * 13 (the input is 416 * 416) scale suitable for detecting large objects was deleted, and only the two detection heads were retained.

3.3.2. Deformable convolution

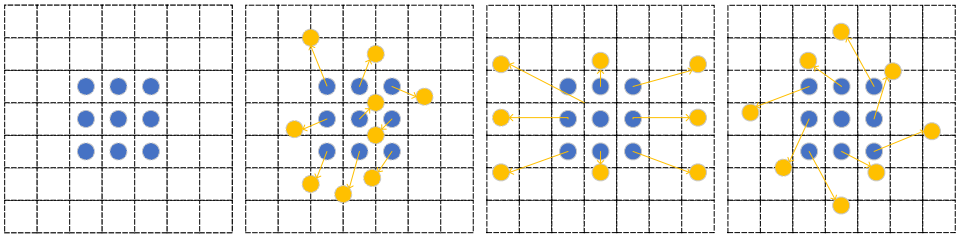
The regular grid R is used by conventional convolution to sample in the feature map and the convolution operation is shown in Figure 4a. The formula of sampling point k_0 in a conventional convolution operation is as follows:

$$y(k_0) = \sum_{k_n \in R} \omega(k_n) \cdot x(k_0 + k_n) \tag{1}$$

An offset to each sampling point is added in the deformable convolution, as shown in Formula (2). At this time, Figure 4a becomes an irregular sampling point in Figure 4b. Figure 4c,d are special cases of the deformable convolution, indicating that the deformable convolution is suitable for various morphological changes. The deformable convolution sampling point k_0 is changed from Formula (1) to Formula (3), where $\omega(k_n)$ is the weight, the offset Δk_n is usually not an integer, and Formula (3) is calculated by bilinear interpolation [29].

$$\{\Delta k_n | n = 1, \ldots, N\}, N \in |R| \tag{2}$$

$$y(k_0) = \sum_{k_n \in R} \omega(k_n) \cdot x(k_0 + k_n + \Delta k_n) \tag{3}$$



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

Figure 4. Comparison of conventional convolution and deformable convolution: (**a**) conventional convolution; (**b**) deformable convolution; (**c**) special case 1; (**d**) special case 2.

Furthermore, the channel number of output characteristics of ordinary convolution operation is N. The same operational deformable convolution can obtain output features with 2N channels. Deformable convolutions have offset features not found in ordinary convolutions. The learning process of a deformable convolution is shown in Figure 5. The offset is obtained by a convolution layer. The convolution kernel in this convolution layer is the same as the ordinary convolution kernel. The output offset feature size is consistent with the input feature map size. The channels of 2N dimension include output characteristics and migration characteristics. During training, the convolution kernel used to generate the output feature and the convolution kernel used to generate the offset are synchronously learning.

The ordinary convolution only samples the fixed position of the input feature map. This convolution geometric transformation modeling ability is weak. Deformable convolution adds two-dimensional migration to the conventional network sampling position in the standard convolution, which can make the grid deform freely and concentrate on the regions or objects of interest. The receptive field changes with the change of features, and the similar structure information adjacent in each pixel is adaptively fused. The deformable convolution is more favorable to the characteristic extraction, thus improving the accuracy of the detection.

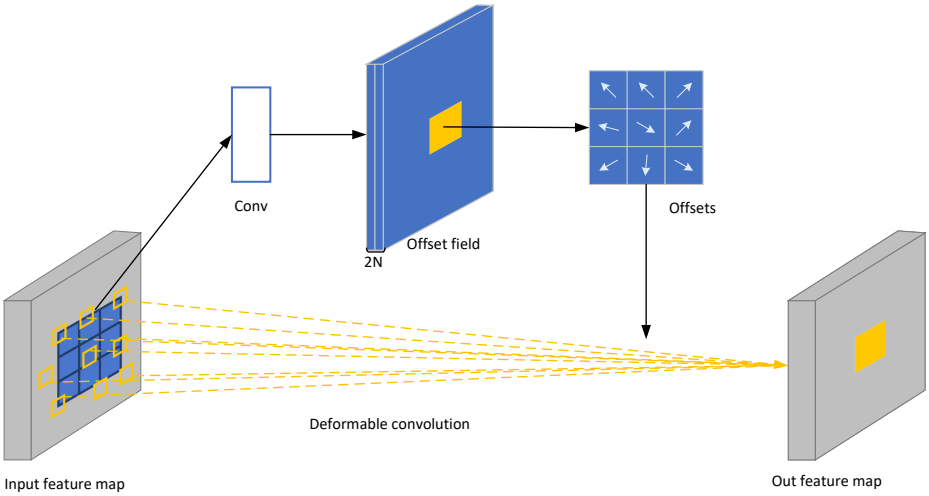


Figure 5. Deformable convolution process.

Three 3 * 3 deformable convolutions were used in the res-dcn [30] module for feature extraction. Then, the output and input were spliced according to the number of channels to retain the multilevel semantic information. The structure of the res-dcn is shown in Figure 6a. Deformable conv consists of a deformable convolution, a BN layer and a Mish activation function, concat represents the connection by the number of channels. The original convolution block structure of N8, N13, N18, N21, and N24 in Figure 2 is shown in Figure 6b. The improved structure CRD is shown in Figure 6c. The original 1 * 1 convolution of the first layer was retained for channel compression. The second-, third-, and fourth-layer convolutions below were replaced by the res-dcn module. A deformable convolution was used to extract more representative vehicle logo features, and the final 1 * 1 convolution was retained to reduce the channel after a concat operation.

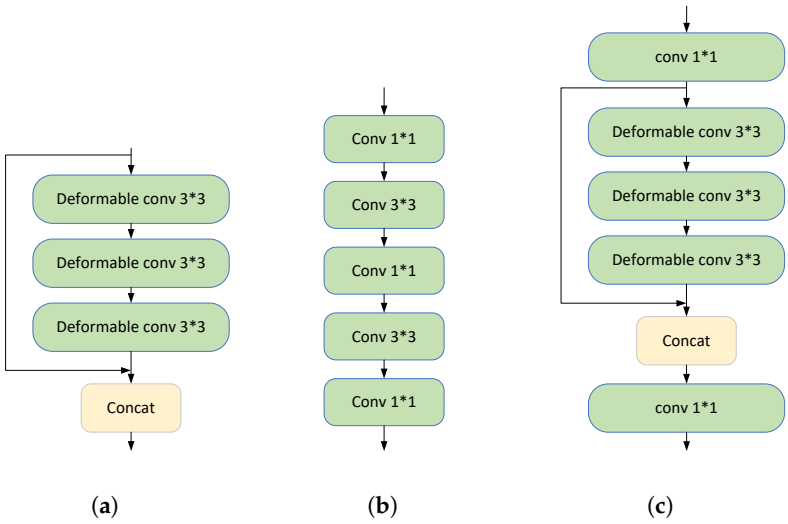


Figure 6. Comparison of original conv module and CRD module: (**a**) res-dcn; (**b**) original conv module; (**c**) CRD module.

Given the irregular shape of vehicle logos, a deformable convolution was introduced in the neck part. The receptive field can change with the change of features, and the deformable feature image was generated to improve the detection accuracy of the vehicle logo.

3.4. CT-Head

Vision Transformer [31] is the visual version of Transformer [32], breaking the isolation between natural language processing (NLP) and computer vision (CV). The advantages of a CNN are scale and distortion invariance. Transformer's strengths are dynamic attention, global context, and a better generalization. CvT [33] introduced a convolution into Transformer, which introduced the ideal properties of a CNN into the VIT architecture, while maintaining the advantages of Transformer. Therefore, in order to give full play to the advantages of a CNN and Transformer, inspired by the paper [34], a convolutional transformer block was introduced into YOLOHead, then CT-Head was proposed. Figure 7 shows the structure of the convolutional transformer block, which consists of two parts. The first part is the multiple-attention layer, and the second part is the full-connection layer.

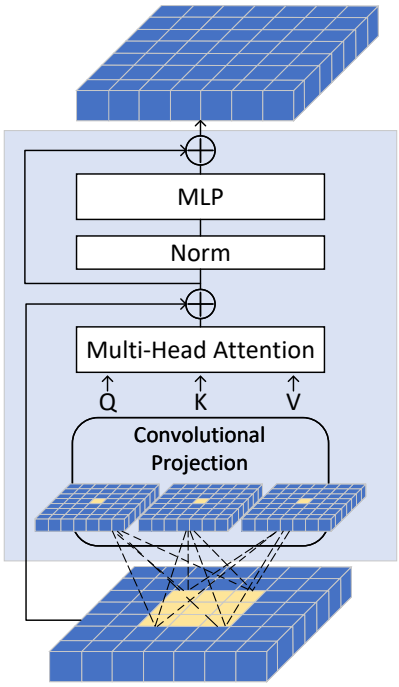


Figure 7. Convolutional transformer block.

The resolution of the feature map at the end of the network is low. The application of the convolutional transformer block on the low-resolution feature map can reduce the cost of calculation and storage. As a result, in YOLOv4, the convolutional transformer block was applied to the head. Feature maps were obtained by the backbone feature extraction network and a series of convolutions in the neck structure. Firstly, the feature map was input into the convolutional transformer block, and Q, K, and V were obtained through a deep separable convolution operation [35], namely a convolution projection. Then, different local information was captured through multiple attention layers and an MLP. The potential of feature representation with self-attention was explored to extract more global information. Finally, the obtained output was used to transform the feature dimension and output through 1 * 1 and 3 * 3 convolution structures. The deeper the network level is, the more obvious the gradient disappears, resulting in the disappearance of small object features. An insufficient utilization of features causes a degradation of detection results. To solve this problem, the residual structure was introduced between the convolutional transformer block and the 3 * 3 convolution. This structure adopted the short-circuit connection mechanism, which can enhance the reuse rate of features.

Therefore, CT-Head could make the transmission of data information in the network more diverse and stronger and enhance the reuse rate of small object features. This structure also combined more global information, which helped to improve the transmission ability of the network to improve the accuracy of vehicle logo detection.

4. Results

4.1. Experimental Environment and Parameter Description

All models in this paper were implemented on the deep learning algorithm framework Pytorch 1.7.1. A GeForce RTX 3090 GPU was used for training and testing in all of our experiments. The input image size of all experiments was 416 * 416. All ablation experiments were trained using the standard stochastic gradient descent method. The initial learning rate was 0.001 and the weight attenuation was 0.0005. The batch size was set to 2, with 150 epochs per experiment.

4.2. Evaluation Metrics

The IoU threshold was set from 0.5 to 0.95, as the threshold to judge whether the object detection was the foreground or background. The evaluation indexes average precision (AP), recall rate (Recall), and mean average precision (mAP) were used to evaluate the experimental results. AP is the mean value of accuracy when IoU is 0.5~0.95 and Recall is 0~1 under the current category. The calculation method of AP is shown in Formula (4). In other words, AP is the area under the two-dimensional curve drawn with the Recall as the horizontal axis and the precision as the vertical axis. MAP is the mean value of the average precision AP for all categories, as shown in Formula (5).

$$AP = \int_0^1 P(r)dr \tag{4}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{5}$$

4.3. Dataset

VLD-45 [7], a new vehicle logo dataset for detection and recognition, was used as the experimental dataset. The dataset is based on the Web crawler technology and Pascal VOC dataset, containing 45 categories with 45,000 images and 50,359 objects.

The maximum image size is 7359 * 4422, with a minimum image size of 610 * 378. The proportion of vehicle signs in the image is only 0.2. Many types of vehicles are present in the VLD-45 dataset, including most of the common vehicle brands in the current market. The dataset presents several research challenges, involving small objects, background interference, shape deformation, low contrast, and other issues. It has very important research value for small object detection tasks; some images in the dataset are shown in Figure 8.

By analyzing the VLD-45 dataset, it was found that some images not only annotated the logo on the vehicle, but also annotated the logo in the background of the vehicle. The purpose of our experiment was to detect and identify the vehicle logo to assist in vehicle information detection. Consequently, we deleted the images that annotated the vehicle logo in the background. A total of 30,000 images were randomly selected from the remaining images, then the images of each category were randomly split into a training set and a test set. The ratio was 1:1, meaning that the training set and the test set each contained 15,000 images separately.

Figure 8. Examples from the dataset.

4.4. Experimental Results and Analysis

The process of the vehicle logo detection experiment is shown in Figure 9. Firstly, the network was used to train the training set images in the data set to obtain the trained model. Then, the test set image was input into the model to detect the vehicle logo and output the experimental results.

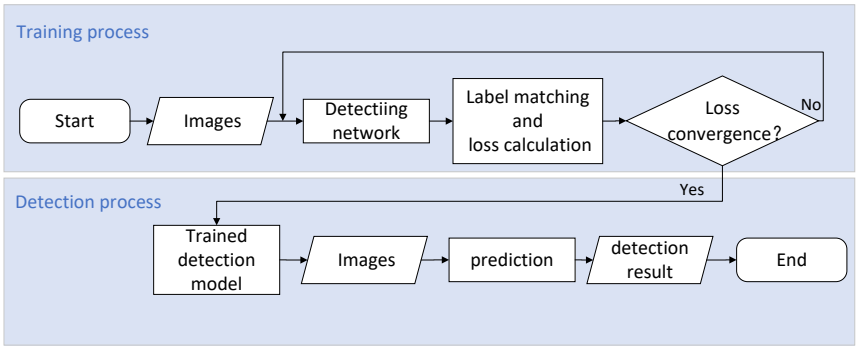


Figure 9. Vehicle logo detection process.

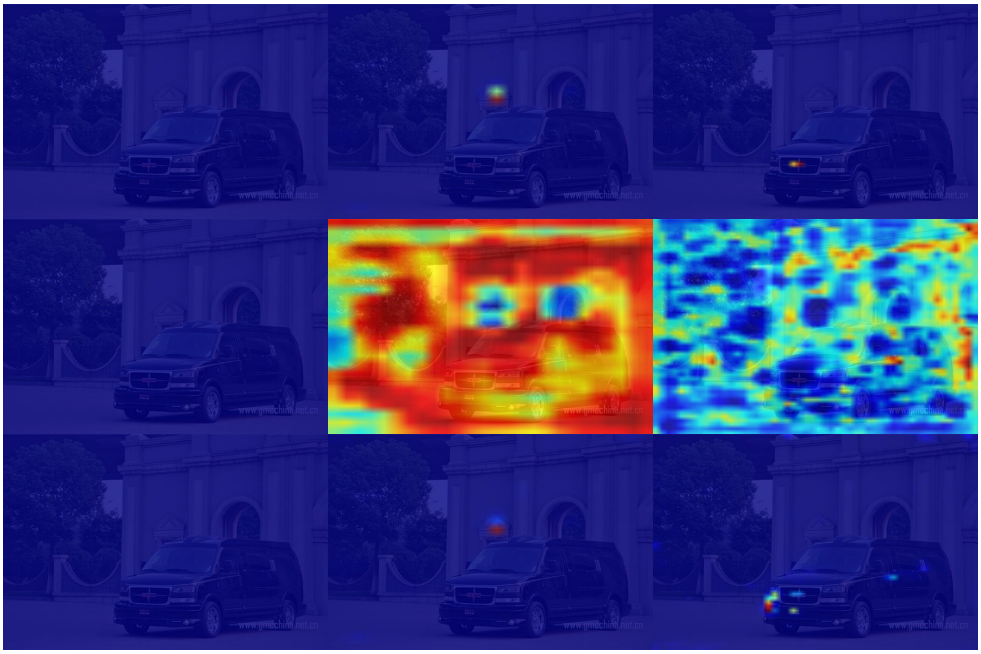
To improve the detection effect of vehicle logo detection, the new output layer, CSP-DenseNet structure, deformable convolution structure, CT-Head, and ECA-Net were introduced into YOLOv4. The above methods were combined with YOLOv4 and ablation experiments were performed on the VLD-45 dataset. The influence of these methods on the performance of vehicle logo detection was quantified by experimental results. The structure and results of each model in the ablation experiment are shown in Table 1. NSF refers to the new structure formed after adding the shallow output layer. DN refers to the residual block composed of the CSPDenseNet module, and RD refers to the res-dcn block. The '✓' in the table indicates that the corresponding module is referenced in the experiment.

Table 1. Structure and results of ablation experimental model.

Experiment	Model	NSF	DN	RD	CT-Head	ECA	mAP (%)	Recall (%)
1	YOLOv4	-	-	-	-	-	57.22	62.05
2	YOLO-NSF	✓	-	-	-	-	58.13	63.11
3	YOLO-NSF-DN	✓	✓	-	-	-	59.70	64.01
4	YOLO-NSF-RD	✓	-	✓	-	-	59.39	60.67
5	CT-YOLO-NSF	✓	-	-	✓	-	59.55	64.49
6	CT-YOLO-NSF-DN-RD	✓	✓	✓	✓	-	61.97	66.30
7	My Model	✓	✓	✓	✓	✓	62.94	67.63

In Table 1, the mAP and Recall values of YOLO-NSF (Experiment 2, NSF module combined with YOLOv4 and a detection head is removed) are all higher than those of YOLOv4 (Experiment 1). The comparative experiment proved that the model could extract more small object features in the vehicle logo detection task after adding the shallow output layer. The region of interest of the model to the detected image can be seen intuitively by a thermodynamic diagram. Figure 10a is the thermodynamic diagram of the 13 * 13 output layer, Figure 10b is the thermodynamic diagram of the 26 * 26 output layer, and Figure 10c is the thermodynamic diagram of the 52 * 52 output layer. Obviously, the 13 * 13 output layer did not pay attention to the characteristics of the vehicle logo, so the improved structure removed this detection head. The results of YOLO-NSF-DN (Experiment 3, DN module combined with YOLO-NSF) and YOLO-NSF comparison experiments showed that the CSPDenseNet introduced in the model significantly improved the accuracy of the vehicle logo detection. The CSPDenseNet structure could improve the reuse rate of features and reduce the impact of small targets on the detection effect. The mAP value of YOLO-NSF-RD (Experiment 4, RD module combined with YOLO-NSF) was higher than that of YOLO-NSF, but the Recall value decreased. The reason was as follows: The deformable convolution was introduced to make the model recognize the complex shape of vehicle logos, which increased the recognition accuracy. However, after adding the deformable convolution, for some pictures with low pixels, the model predicted some positive classes into negative classes, resulting in a decrease in the Recall value. CT-YOLO-NSF (Experiment 5) replaced the head of YOLO-NSF with the proposed CT-Head. Compared with the results of YOLO-NSF, the mAP and Recall values increased by more than 1%. The experiment proved that after CT-Head was applied to the model, the model could focus on both local and global features, reducing the impact of the background on the vehicle logo detection. After these improvements, in order to further make the network pay more attention to the characteristics of vehicle logos, ECA-Net was introduced to further improve the detection accuracy of vehicle logos. My Model (Experiment 7) was an improved model that integrated all the above methods. From the results, the detection effect was greatly improved.

From the model recognition effect, the YOLOv4 model was not ideal for vehicle logo detection. The main factor was that the images in the VLD-45 dataset were small objects, and the object background was complex. YOLOv4 had nice detection performance for multiscale objects and simple background information of vehicle logo images. However, it had a poor detection effect for images with a similar texture and background to the vehicle logo. Figures 11 and 12 show the performance of YOLOv4 and improved YOLOv4 in some images from the VLD-45 dataset.



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

Figure 10. Comparison of heatmaps of YOLOv4 output layers: (**a**) 13 * 13 output layer; (**b**) 26 * 26 output layer; (**c**) 52 * 52 output layer.

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

Figure 11. Comparison 1 and comparison 2 of the effect of YOLOv4 and My Model: (**a**) example 1, YOLOv4 detection results; (**b**) example 1, My Model detection results; (**c**) example 2, YOLOv4 detection results; (**d**) example 2, My model detection results.

Figure 11a shows the test result of YOLOv4, and Figure 11b shows the test result of the improved model. YOLOv4 identified some background information incorrectly as a vehicle logo, and the improved model reduced this error. Figure 11c shows the detection results of YOLOv4 with a certain degree of missed detection. Figure 11d shows the test result of the improved model, reducing the missing rate of the vehicle mark.

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

Figure 12. Comparison 3 and comparison 4 of the effect of YOLOv4 and My Model: (**a**) example 3, YOLOv4 detection results; (**b**) example 3, My Model detection results; (**c**) example 4, YOLOv4 detection results; (**d**) example 4, My model detection results.

It can be seen from Figure 12a that for some vehicle images in natural scenes, YOLOv4 could not successfully label the vehicle logo position. The improved model could successfully detect the vehicle logo and accurately label the position and had a high confidence, as shown in Figure 12b. Figure 12c,d show that YOLOv4 had a certain degree of false detection, and the false detection rate was reduced after improvement. It can be seen from these test results that our model had a better detection effect than YOLOv4.

Table 2 shows all kinds of AP values detected by YOLOv4 and the improved YOLOv4 model (My Model) on the VLD-45 dataset (the average values of APs with IOUs from 0.5 to 0.95). The original YOLOv4 model showed good performance in the detection of 45 categories in the dataset. However, the false detection rate still needed to be lower, especially the LINCOLN logo detection. Although the logo shape of this class was not complex, it was difficult to distinguish from background information, resulting in a low AP value for this class. As can be seen from the results of Table 2, compared with YOLOv4, in the improved model, all classes' AP values were increased. The LINCOLN class's AP value was significantly increased.

To further verify the effectiveness of the proposed improved model (My Model), the commonly used object detection models SSD, RetinaNet [36], Free Anchor [37], Centernet [38], EfficientDet [39], and YOLOv5 [40] were compared and tested. The experimental results are shown in Table 3. The table lists the mAP (IoU from 0.5 to 0.95), AP_{50} (IoU = 0.5), AP_{75} (IoU = 0.75), AP_S (mAP for small objects), and Recall of all categories detected by each model on the VLD-45 dataset. The values of all the evaluation indicators of our model were higher. Especially for the AP_S value of small object detection results, our model was much higher than the other models in the table. This shows that our model was more suitable for small object detection. The detection effects of these models on the dataset VLD-45 are shown in Figure 13. The improved YOLOv4 model performed better than the other models in Table 3.

Overall, SSD and RetinaNet had poor detection results due to fewer parameters and insufficient feature extraction. Other mainstream models performed well on the VLD-45 dataset. However, due to the difficulty of vehicle logo detection, the detection accuracy needed to be further improved. The model proposed in this paper fully considered the characteristics of the object in the dataset. Aiming at the problems of small objects, complex object backgrounds, and irregular shapes in the dataset, the model was improved. Therefore, the detection effect on the VLD-45 dataset was better than that of the above mainstream models, and it also showed good performance in various types of detection.

Table 2. AP values of YOLOv4 and improved YOLOv4 in various categories.

Number	Class	AP (YOLOv4)	AP (My Model)
01	BAIC Group	0.67	0.70 (+0.03)
02	Ford	0.56	0.63 (+0.07)
03	SKODA	0.51	0.59 (+0.08)
04	Venucia	0.59	0.66 (+0.07)
05	HONDA	0.63	0.66 (+0.03)
06	NISSAN	0.59	0.67 (+0.08)
07	Cadillac	0.62	0.67 (+0.05)
08	SUZUKI	0.59	0.62 (+0.03)
09	GEELY	0.55	0.62 (+0.07)
10	Porsche	0.48	0.51 (+0.03)
11	Jeep	0.45	0.52 (+0.07)
12	BAOJUN	0.58	0.67 (+0.09)
13	ROEWE	0.62	0.67 (+0.05)
14	LINCOLN	0.40	0.52 (+0.12)
15	TOYOTA	0.60	0.67 (+0.07)
16	Buick	0.60	0.67 (+0.07)
17	CHERY	0.52	0.57 (+0.05)
18	KIA	0.59	0.62 (+0.03)
19	HAVAL	0.47	0.48 (+0.01)
20	Audi	0.65	0.68 (+0.03)
21	LAND ROVER	0.55	0.56 (+0.01)
22	Volkswagen	0.59	0.70 (+0.11)
23	Trumpchi	0.60	0.65 (+0.05)
24	CHANGAN	0.54	0.61 (+0.07)
25	Morris Garages	0.62	0.69 (+0.07)
26	Renault	0.54	0.64 (+0.10)
27	LEXUS	0.67	0.71 (+0.04)
28	BMW	0.57	0.62 (+0.05)
29	MAZDA	0.64	0.70 (+0.06)
30	Mercedes-Benz	0.63	0.73 (+0.10)
31	HYUNDAI	0.58	0.64 (+0.06)
32	Chevrolet	0.53	0.58 (+0.05)
33	BYD	0.61	0.63 (+0.02)
34	PEUGEOT	0.54	0.59 (+0.05)
35	Citroen	0.56	0.62 (+0.06)
36	Brilliance Auto	0.60	0.63 (+0.03)
37	Volvo	0.57	0.62 (+0.05)
38	Mitsubishi	0.52	0.57 (+0.05)
39	Subaru	0.55	0.59 (+0.04)
40	GMC	0.53	0.57 (+0.04)
41	Infiniti	0.60	0.66 (+0.06)
42	FAW Haima	0.61	0.67 (+0.06)
43	SGMW	0.62	0.65 (+0.03)
44	Soueast Motor	0.60	0.67 (+0.07)
45	QOROS	0.54	0.62 (+0.08)

(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**)

Figure 13. Comparison of detection effect of each detection model: (**a**) example of detection effect of SSD; (**b**) example of detection effect of RetinaNet; (**c**) example of detection effect of Free Anchor; (**d**) example of detection effect of Centernet; (**e**) example of detection effect of EfficientDet_d2; (**f**) example of detection effect of YOLOv5; (**g**) example of detection effect of My Model.

Table 3. Comparison results of mainstream detection models.

Model	Backbone	mAP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	Recall (%)
SSD	VGG	49.87	85.5	52.8	41.2	51.64
RetinaNet	Resnet50	44.33	65.4	54.8	31.1	40.12
Free Anchor	Resnet50	56.20	92.7	63.6	50.7	63.70
Centernet	Resnet50	56.37	93.1	63.8	49.0	54.68
EfficientDet_d2	Efficient	59.00	86.6	72.4	51.8	63.10
YOLOv5	CSPDarknet	56.46	94.2	62.4	48.3	57.64
My Model	CSPVLnet	62.94	98.5	76.3	58.9	67.63

5. Discussion

In the literature, manual features were first introduced and used for vehicle logo detection. These features achieved good results in detection. However, they could not achieve the expected detection performance on vehicle logos in a real environment. The vehicle logo detection method based on deep learning was introduced later. The method based on deep learning could detect vehicle logos well in some natural scenes. However, due to the small target and complex background information, the detection accuracy of vehicle logos still needed to be improved. Our work selected a larger vehicle logo detection dataset, VLD-45. We improved the vehicle logo detection effect. However, to reduce the impact of the vehicle logo shape on the detection results, we introduced a deformable convolution in the improvement process. This led to a decrease in our detection speed. In future work, we will continue to improve the part that introduces the deformable

convolution. We will improve the structure of this part, reduce the number of parameters, and improve the vehicle logo detection speed.

6. Conclusions

In this paper, to solve the problem of low recognition rate caused by small objects, multiple types, and a complex background around vehicle logos, an improved YOLOv4-based vehicle logo recognition method was presented. To integrate more shallow information, a shallow output layer was added to change the neck structure of the original model. The combination of shallow location information and deep semantic information reduced the loss of small object features in vehicle logo recognition. To improve the reuse rate of small object features, the CSPDenseNet module was integrated into Darknet53. Moreover, to make the model more accurately fit the shape and position characteristics of different vehicle logos, a deformable convolution and the ECA-Net were integrated into the neck part. Finally, to capture richer context information and more global information and achieve a more accurate positioning of vehicle logos in different backgrounds, the convolutional transformer block was introduced into the detection head part of the model. Compared with the original model on the VLD-45 dataset, the mAP (IoU from 0.5 to 0.95) of the improved vehicle logo recognition model was increased by 5.72%, and the Recall was increased by 5.58%.

Author Contributions: Writing—original draft preparation, X.J.; formal analysis, K.S.; data curation, L.M.; writing—review and editing, Z.Q.; project administration, C.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Youth Innovation Team Development Plan of Shandong Province Higher Education grant number 2019KJN048.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study did not report any data.

Acknowledgments: We thank Shuo Yang's research team for providing us with their dataset, VLD-45, which was a great help to our experiment. This work was supported by the Youth Innovation Team Development Plan of Shandong Province Higher Education (2019KJN048).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Raskar, R.R.; Dabhade, R.G. Automatic Number Plate Recognition (ANPR). Available online: https://www.scinapse.io/papers/2320385793 (accessed on 15 August 2022).
2. Llorca, D.F.; Arroyo, R.; Sotelo, M. Vehicle logo recognition in traffic images using HOG features and SVM. In Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC), Hague, The Netherlands, 6–9 October 2013.
3. Soon, F.C.; Hui, Y.K.; Chuah, J.H. Pattern recognition of Vehicle Logo using Tchebichef and Legendre moment. In Proceedings of the 2015 IEEE Student Conference on Research and Development (SCOReD), Kuala Lumpur, Malaysia, 13–14 December 2015.
4. Yu, S.; Zheng, S.; Hua, Y.; Liang, L. Vehicle logo recognition based on Bag-of-Words. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, Krakow, Poland, 27–30 August 2013.
5. Huang, Y.; Wu, R.; Sun, Y.; Wang, W.; Ding, X. Vehicle Logo Recognition System Based on Convolutional Neural Networks with a Pretraining Strategy. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 1951–1960. [CrossRef]
6. Yu, Y.; Li, H.; Wang, J.; Min, H.; Chen, C. A Multilayer Pyramid Network Based on Learning for Vehicle Logo Recognition. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 3123–3134. [CrossRef]
7. Yang, S.; Bo, C.; Zhang, J.; Gao, P.; Li, Y.; Serikawa, S. VLD-45: A big dataset for vehicle logo recognition and detection. *IEEE Trans. Intell. Transp. Syst.* **2021**, 1–7. [CrossRef]
8. Zhang, J.; Chen, L.; Bo, C.; Yang, S. Multi-Scale Vehicle Logo Detector. *Mob. Netw. Appl.* **2021**, *26*, 67–76. [CrossRef]
9. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
10. Pan, H.; Zhang, B. An integrative approach to accurate vehicle logo detection. *J. Electr. Comput. Eng.* **2013**, *2013*, 18. [CrossRef]

11. Thubsaeng, W.; Kawewong, A.; Patanukhom, K. Vehicle logo detection using convolutional neural network and pyramid of histogram of oriented gradients. In Proceedings of the 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), Chon Buri, Thailand, 14–16 May 2014; pp. 34–39.
12. Peng, H.; Wang, X.; Wang, H.; Yang, W. Recognition of low-resolution logos in vehicle images based on statistical random sparse distribution. *IEEE Trans. Intell. Transp. Syst.* **2014**, *16*, 681–691. [CrossRef]
13. Zhao, J.; Wang, X. Vehicle-logo recognition based on modified HU invariant moments and SVM. *Multimed. Tools Appl.* **2019**, *78*, 75–97. [CrossRef]
14. Yu, Y.; Wang, J.; Lu, J.; Xie, Y.; Nie, Z. Vehicle logo recognition based on overlapping enhanced patterns of oriented edge magnitudes. *Comput. Electr. Eng.* **2018**, *71*, 273–283. [CrossRef]
15. Xia, Y.; Jing, F.; Zhang, B. Vehicle Logo Recognition and attributes prediction by multi-task learning with CNN. In Proceedings of the 2016 12th International Conference on Natural Computation and 13th Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, 13–15 August 2016.
16. Li, B.; Hu, X. Effective vehicle logo recognition in real-world application using mapreduce based convolutional neural networks with a pre-training strategy. *J. Intell. Fuzzy Syst.* **2018**, *34*, 1985–1994. [CrossRef]
17. Yu, Y.; Guan, H.; Li, D.; Yu, C. A Cascaded Deep Convolutional Network for Vehicle Logo Recognition From Frontal and Rear Images of Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2019**, *22*, 758–771. [CrossRef]
18. Ke, X.; Du, P. Vehicle logo recognition with small sample problem in complex scene based on data augmentation. *Math. Probl. Eng.* **2020**, *2020*, 6591873. [CrossRef]
19. Liu, R.; Han, Q.; Min, W.; Zhou, L.; Xu, J. Vehicle logo recognition based on enhanced matching for small objects, constrained region and SSFPD network. *Sensors* **2019**, *19*, 4528. [CrossRef]
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
22. Zhang, J.; Yang, S.; Bo, C.; Zhang, Z. Vehicle logo detection based on deep convolutional networks. *Comput. Electr. Eng.* **2021**, *90*, 107004. [CrossRef]
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *arXiv* **2015**, arXiv:1512.02325.
24. Lu, W.; Zhao, H.; He, Q.; Huang, H.; Jin, X. Category-consistent deep network learning for accurate vehicle logo recognition. *Neurocomputing* **2021**, *463*, 623–636. [CrossRef]
25. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D. Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors. *arXiv* **2020**, arXiv:2008.07043.
26. Zand, M.; Etemad, A.; Greenspan, M. Oriented Bounding Boxes for Small and Freely Rotated Objects. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–15. [CrossRef]
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* 2019, arXiv:1910.03151.
28. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
29. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *arXiv* **2015**, arXiv:1506.02025.
30. Q, H.F.; M, C.; F, F.G. Improved YOLO object detection algorithm based on deformable convolution. *Comput. Eng.* **2021**, *47*, 8.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762v5.
33. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 22–31.
34. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
35. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
36. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *1*, 2999–3007.
37. Zhang, X.; Wan, F.; Liu, C.; Ji, R.; Ye, Q. FreeAnchor: Learning to Match Anchors for Visual Object Detection. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 147–155.
38. Zhou, X.; Wang, D.; Krhenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.

39. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WD, USA, 14–19 June 2020.
40. Jocher, G.; Stoken, A.; Borovec, J.; Chaurasia, A.; Changyu, L.; Laughing, A.; Hogan, A.; Hajek, J.; Diaconu, L.; Marc, Y.; et al. ultralytics/yolov5: v5.0-YOLOv5-P6 1280 models AWS Supervise. ly and YouTube integrations. *Zenodo* **2021**, *11*. Available online: https://zenodo.org/record/4679653#.Y1ENIXZBxPY (accessed on 15 August 2022).