Vehicle Logo Detection Method Based on Improved YOLOX

Ma Cheng[1], Jiang Xiaoli[1], Qu Zhijian[1*]
[1](*School of Computer Science and Technology, Shandong University of Technology, China)*

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 YOLOX model was presented. Firstly, the Swin Transformer structure is introduced to replace part of the CSP structure to improve the backbone feature extraction network to make the model better. Secondly, in order to highlight the local key features and prevent the loss of vehicle logo information, combined with asymmetric convolution, a CSPAC structure is proposed to replace the original CSP structure of the network neck. Finally, a Bicat feature fusion method is proposed to enhance the detection model 's learning of channel features with stronger expression ability, and focus on vehicle logo information with high-weight channels to strengthen the network 's learning of vehicle logo features. Experimental results showed that the average accuracy of all categories in the VLD-45 dataset was 64.21%, which was 5.86% higher than the original model. It indicated that the improved model could perform well in vehicle logo detection.*

KEYWORDS *-Vehicle logo detection, Small object detection, Transformer, Asymmetric Convolution, Attention mechanism*

I. INTRODUCTION

With the development of science and technology, intelligent transportation system has become an important part of people 's life. 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 diffificult to replace, so its recognition is of great signifificance. Vehicle identification plays an important role in maintaining traffic safety and security management. Vehicle logo is a key symbol of automobile manufacturers. It carries important information of vehicle identification and is difficult to forge. Therefore, accurate identification of vehicle logos helps to supplement vehicle information and plays an important role in traffic accidents and automobile-related crimes.

The existing vehicle sign recognition methods are mainly divided into two categories, based on traditional machine learning methods and based on depth. Traditional machine learning methods train detectors by designing different hand-crafted features. However, the specially designed detector is not enough to complete the identification of so many types of vehicle logos in a complex natural environment. In recent years, object detection methods based on deep learning have dominated computer vision tasks. Some deep learning frameworks have been applied to vehicle sign detection.

Although the current deep learning model has achieved some results in the field of vehicle logo recognition, there are still some problems in its recognition and classification. The vehicle logo makes up only0.1% to 1% of the overall image. The small size has a signifificant 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 shape of the vehicle logo is different, and the irregular shape brings great difficulty to the detection. These factors lead to the low accuracy of the existing vehicle sign target detection method based on deep learning. Therefore, a target detection method that can effectively identify vehicle signs in the natural environment is needed.

To solve the above problems, a vehicle logo recognition method based on the improved YOLOX[1] 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: 1.Swin Transformer Block is introduced to enhance the network 's ability to focus on global features and reduce the interference of background information on vehicle sign detection tasks. 2.The CSPAC module is designed to improve the convolution method of the neck, expand the receptive field in the process of strengthening feature extraction and make the network pay more attention to the key features of the vehicle logo, reduce the omission of features, and improve the low detection accuracy caused by the complexity and diversity of the vehicle logo. 3.Combined with BiFPN, a Bicat feature fusion method is proposed, which is equivalent to using the attention mechanism in the network to solve the problem that the logo occupies fewer pixels and is a small target.

II. RELATED WORK

The current vehicle sign detection methods with better results include traditional machine learning methods and deep learning methods. In the early stage of vehicle logo detection research, most of the traditional manual feature extraction methods are used for vehicle logo detection. Firstly, the vehicle logo is described by histogram, texture, invariant moment and other features, and then the machine learning algorithm is used to identify and detect the vehicle logo.Pan[2] proposed a fast and reliable vehicle logo detection method, which combines Haar-like features(Haar) features and Histograms of Oriented Gradients(HOG)[3]. Thubsaeng[4] proposed a new method for detecting and recognizing vehicle signs from front and rear view images of vehicles. This method is a two-stage method that combines Convolutional Neural Network(CNN) and pyramid histogram of oriented gradients(PHOG) features. Zhao[5] proposed to increase the invariant moment feature to improve the recognition effect for the problem of vehicle logo recognition under weak light conditions. Peng[6] proposed a new feature representation strategy, strategy random sparse distribution(SRSD), for low-resolution and low-quality images collected in intelligent transportat systems. Yu[7] constructed a new vehicle logo recognition dataset(HFUT-VL), and proposed a new vehicle logo recognition method based on this dataset. The vehicle logo detection method based on traditional features is simple and efficient, but there are few types of vehicle logos that can be detected. There is no way to accurately detect and identify most of the common types of vehicle logos in daily life. On the other hand, the method of artificially extracting vehicle logo features has low mobility and insufficient robustness, which is difficult to meet the actual application requirements.

The rapid development of deep learning provides a new idea for the task of automatic and accurate recognition of vehicle signs. The application of deep learning model to vehicle logo detection has achieved good results. The vehicle logo detection method based on deep learning does not need to manually design features, but automatically learns vehicle logo feature expression from image data.

Some studies have chosen to improve the accuracy of vehicle sign detection from the pre-training strategy. Huang[8] migrated the convolutional neural network to the vehicle logo recognition task, and introduced a pre-training strategy to achieve breakthrough detection results on a large-scale 10-category data set. Li and Hu[9] also proposed a vehicle logo detection method with pre-training strategy. They used the Hadoop framework to preprocess the data and trained a deep convolutional neural network model for logo detection. However, this improved method has a high dependence on the data set, and it is difficult to solve the problem of vehicle logo detection in essence.

In recent years, the research on vehicle logo detection is mainly based on improving the deep learning network structure. Scholars have proposed a vehicle logo detection network framework with higher and higher detection accuracy. Yu[10] proposed a two-stage vehicle logo recognition and detection framework based on cascaded deep convolutional neural networks. This method does not need to rely on the detection of the license plate to locate the vehicle logo, but can directly detect the vehicle logo, which solves the problem of machine learning method to detect the vehicle logo. Yu[11] proposed a Multilayer Pyramid Network Based on Learning(MLPNL). The network considers multiple resolutions to extract valuable features, thereby improving the performance of the model in vehicle logo detection tasks. Liu[12] proposed a vehicle logo recognition method based on enhanced matching, constrained region extraction and SSFPD(single-shot feature pyramid detector) network. Firstly, the method detects the constrained area based on Faster-RCNN and extracts the information of the front and rear of the vehicle. Then, in the training process, data enhancement is performed by copying and pasting the vehicle logo multiple times in the constrained area. Finally, based on the Res Next network, the SSFPD network is improved to extract features and generate feature maps. The literature method improves the accuracy of small target detection by improving the characteristics of small target. However, there are also some limitations. The data set used in the experiment has only 13 types of vehicle logos, and the applicable scenarios are limited. Lu[13] proposed a new category-consistent deep learning network framework for accurate vehicle logo detection and recognition. The model framework proposed in this paper mainly includes two parts. The first part is a new feature extraction network(VLF-net) to extract vehicle logo features, which extracts hierarchical features by considering the high-level and low-level features of the image. The second part is a new category-consistent mask-learning module(CCML), which helps the framework focus on type-consistent regions. The proposed category consistent mask learning module can learn the category area without knowing the accurate vehicle logo area, so that the network can focus on this area. The detection framework proposed in this paper has shown good performance on several public datasets. The outstanding contribution is to avoid license plate detection and no longer rely on the labeling of artificial bounding boxes. However, this method can only identify the sign from the front image of the vehicle, which has high requirements for the quality of the input image, and the limitation of the detection task is large.

From the above results of using deep learning methods to study vehicle sign target detection, it is not difficult to find that the type and number of vehicle images in the data set used also play an important role in the research process. Therefore, Xia[14] extended the vehicle logo recognition data set of Xiamen University and studied it with the extended data set. They proposed a method that combines CNN with multi-task learning to identify and predict vehicle logos. Ke and Du[15] optimized the data for the problem of small logo area and small number of pictures in the data set, and proposed three logo data enhancement strategies. The cross sliding segmentation method is used for the problem of small data sample size. In order to expand the area of the logo in the image, a small border method is proposed. A vehicle logo segmentation method based on Gaussian distribution is proposed to enrich the position difference of vehicle logo in the image. These optimization methods reflect the characteristics of the vehicle logo better than the previous traditional methods. They used these methods to optimize the original data set and tested the enhanced data set under some target detection frameworks, and achieved good detection results. Liu[16] proposed a large-scale vehicle logo benchmark dataset(VLD1.0), which contains 66 categories, covering images with different lighting conditions and resolutions. However, the number distribution of each type of vehicle logo image in the VLD1.0 data set is not balanced, which affects the comparative judgment of each type of vehicle logo detection results. Yang[17] constructed a new VLD-30 dataset for vehicle logo detection and recognition, and proposed a fast vehicle logo detection model. Based on the VLD-30 data set, Yang's team[18] continued to expand the data set and enrich the types of vehicle logos in the data set, and proposed a 45-class large vehicle logo data set (VLD-45) for vehicle logo recognition and detection. This paper evaluates the VLD-45 dataset using existing classifiers and detectors. The experimental results show that the data set constructed in this paper has very important research value for small target detection tasks.

In summary, although the method based on deep learning has achieved good results in vehicle logo detection, the accuracy of small target recognition in complex environments still needs to be improved. In recent years, more and more scholars began to study the method of small target detection. Li[19] proposes an adaptive point for the problem of small air targets and chaotic environment. Aiming at the problem of low efficiency of small target detection, Yang[20] proposed a new query mechanism to speed up object detection based on feature pyramid. Akshatha[21] proposed a pedestrian detection dataset(Manipal-UAV) for the problem of small target detection with few public datasets. Zhong[22] proposed a multi-scale contrast enhancement method for infrared small target detection. The detection of vehicle logo belongs to small target detection. The research of vehicle logo detection method is helpful to promote small target detection. The VLD-45 data set is a data set with rich types and a large number of images in the current vehicle logo data set. Most of the vehicle logos in the data set belong to small targets, so this paper will propose a method for the VLD-45 data set.

III. METHODOLOGY

3.1 *Model Overview*

A vehicle logo recognition model based on YOLOX is presented in this paper. YOLOX is an open-source high-performance detector, and its researchers used YOLOv3[23] as the baseline model to propose YOLOX. YOLOX cleverly integrates excellent progress in the field of target detection such as combined decoupling head, data enhancement, and no anchor frame. This model mainly includes three modules: backbone (CSP Darknet), neck (path aggregation network PAN and feature pyramid FPN) and head (YOLO Head).

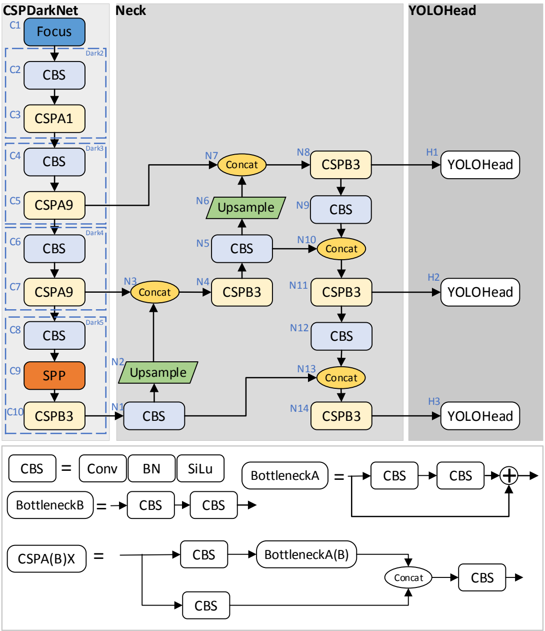

Figure 1. YOLOX Structure

The structure of the YOLOX model is shown in Figure 1. Each part of the model is marked with blue letters and numbers, and the three main parts of the structure of YOLOX are divided. Firstly, the preprocessed image is input into the main feature extraction network CSP DarkNet for feature extraction. Then, the output features of the second, third and fourth CSP[24] structure layers (C5, C7 and C10 in Figure 1) of the feature extraction network are obtained as three-scale output features. Then, the three feature layers obtained are input into the feature fusion network (Neck part in Figure 1) for feature fusion, and the feature extraction is strengthened to obtain the three-scale features after repeated feature fusion and full extraction of information. Finally, the three-scale features are input into the detection layer(YOLOHead part in Figure 1), and the final network weights are obtained by calculating the loss function.

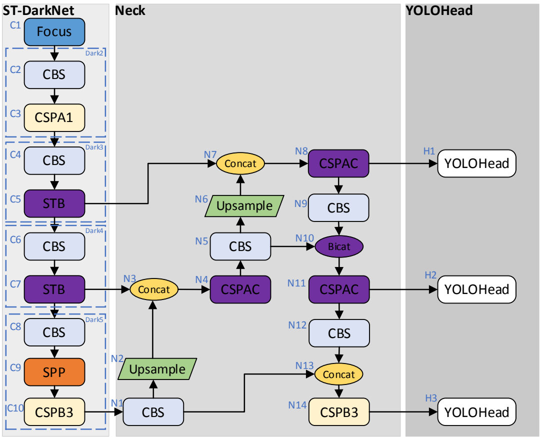

Figure 2. Structure of the improved YOLOX model

According to the characteristics of the vehicle logo detection task, the YOLOX target detection is improved, and the improved model structure is shown in Figure 2. In order to facilitate the introduction of the improved model, each module is marked with blue on the left side. Specifically, the improvement part includes: Swin Transform Block[25] is used instead of C5 and C7; The CSP structure(CSPB3 in the graph) in the neck enhancement feature extraction network is improved to the proposed CSPAC structure ; At N10, the original Concat feature fusion method is no longer used, but a simple attention mechanism is added to the feature channel(Bicat Method).

3.2*ST-Darknet feature extraction network*

The backbone feature extraction network of the original YOLOX is composed of a series of CSP modules and convolutional layers. The ordinary convolution layer is a local operation, which is usually limited to modeling the relationship between adjacent pixels. The continuous stacking of CNN makes the network pay more attention to local information to a certain extent. The shape of the vehicle logo is diverse and complex, and the background information of the vehicle logo in the image is very complex. It is difficult to detect and identify the vehicle logo simply by local features, and it is difficult to achieve the ideal detection effect. Vision Transformer (ViT)[26]can obtain global modeling ability through global self-attention, which can improve the limitations of ordinary convolution. Due to the single size and low resolution characteristics of ViT processing, the characteristics of multi-scale features are weak. From the perspective of computational complexity, the computational complexity of global self-attention is twice the square of image size. Swin Transformer combines the design concepts and prior knowledge of many convolutional neural networks to calculate self-attention within a small window. Therefore, as long as the window size is fixed, the complexity of self-attention calculation is also fixed. Therefore, the total computational complexity is controllable, which reduces the length of the sequence and reduces the computational complexity.

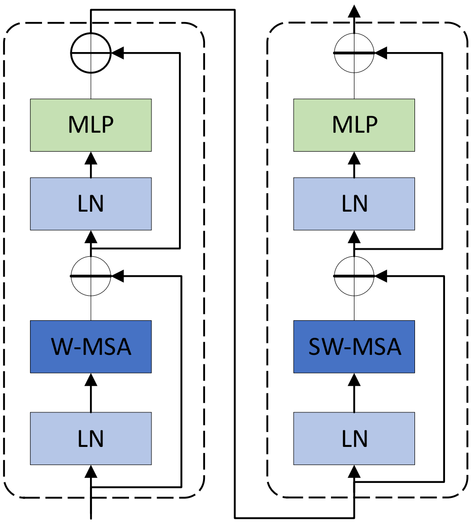

Figure 3. Swin Transformer Block Structure

Therefore, the original backbone feature extraction network is combined with Swin Transformer Block to reconstruct a new feature extraction network ST-Darknet. The feature extraction network thus obtained can also pay attention to local features. The combination of local features and global features can better extract vehicle logo information, which is conducive to the learning of vehicle logo features by the network. The Swin Transformer Block is used to replace the C5 and C7 parts of the YOLOX original backbone feature extraction network to obtain the ST-Darknet network structure. The structure of Swin Transformer Block is shown in Figure 3.

As shown in Figure 3, there are two continuous Swin Transformer Blocks, whose main structures are window multi-headed self-attention (W-MSA) and moving window multi-headed self-attention(SW-MSA).

W-MSA performs Transformer operation in a small window, which can obtain information inside the window. SW-MSA is used to obtain information between windows, including moving window, cyclic shift and mask. Moving window operation can extract information across windows. In order to solve the problem that the nine window sizes after moving the window are inconsistent and cannot be calculated in batches, the cyclic shift method is needed. The cyclic shift is four windows, using mask operation, so that different regions of a window can use a forward to calculate the attention, after calculating the attention, the cyclic shift is restored. ST-DarkNet stacks the above W-MSA and SW-MSA processes after Focus and CSP operations. And constantly adjust the number of channels with convolution, so as to fully extract features in the process of continuous downsampling. At the same time, compared with the original CSP structure used in the structure, the self-attention mechanism in Swin Transformer Block can make the model suppress irrelevant features and better extract effective features, which is conducive to network training.

3.3*CSPAC Structure*

The feature pyramid enhancement feature extraction part of YOLOX is similar to the backbone feature extraction network, and also uses the CSP structure. The two channels of the CSP structure are transformed with a 1*1 convolution respectively, and a 1*1 convolution is used to reduce the dimension after the feature fusion of the two channels. The 1*1 convolution is a commonly used module for lifting dimensions, but the receptive field is relatively small, which is not helpful for extracting features. In addition, the distribution of knowledge learned by the ordinary convolution kernel is uneven, because the center cross position is usually larger on the skeleton part of the convolution kernel. Compared with the weight on the corner, removing the weight on the skeleton is more likely to lead to a decrease in detection accuracy. In general, the model learns to enhance the skeleton part of the convolution in each layer of the network. The Asymmetric Convolion Block(ACB)[27] adds horizontal and vertical kernels to the skeleton to ensure that the skeleton is stronger and follows the properties of the ordinary kernel. Therefore, the ACB module can enhance the square kernel convolution in the horizontal and vertical directions. Highlighting local key features will not cause loss of features while adjusting dimensions. The ACB structure is introduced into the CSP module in the original YOLOX feature pyramid. The ACB structure is used to improve the original conventional 1*1 convolution to obtain the improved CSPAC module to help the model pay more attention to the key features. The structure of CSPAC is shown in Figure 4.

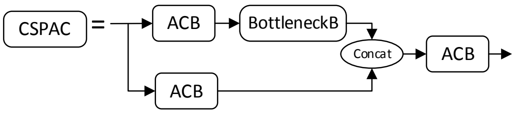

Figure 4.CSPAC Structure

After the fusion of the up-sampled features and the features extracted from the backbone feature extraction network, the 3*3 ACB module is input into the backbone edge and the residual edge respectively. After the ACB module broadens the receptive field rich features, it continues to further extract features through the residual block. The backbone edge continuously increases the convolutional layer through the residual structure to. After learning the features through the residual structure, the main edge and the residual edge are fused. Finally, through the ACB module to re-strengthen the convolution of the skeleton part of the weight of the learning, to improve the detection accuracy.

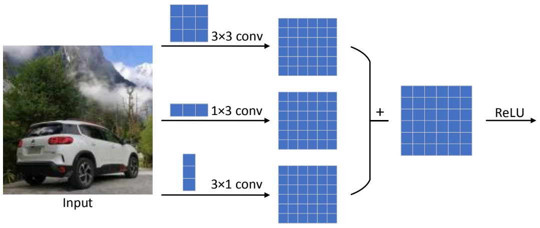

Figure 5.ACB Structure

The ACB is an innovative structure that can replace the standard convolutional layer with a square kernel, as shown in Figure5. The ACB contains three kernels, which are three layers of 3*3,1*3 and 3*1 convolution kernels, and the output of the three layers is summed to enrich the feature space. Similar to the common practice in standard CNN, each layer in the three layers of the ACB module is batch normalized, called a branch, and the output of the three branches is added to the output of the ACB.

$$y=\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}F_{h,w,c}^{(j)}X_{h,w,c} \quad (1)$$

As shown in Figure 6, the sliding window can intuitively explain the additivity of 2D convolutions with different size convolution kernels. The three convolutional layers use the same input, and only the sliding window is depicted in the upper left and lower right corners. The key to maintaining additivity is that the three layers share the same sliding window. Therefore, by adding the kernels of Conv2 and Conv3 to the corresponding position of Conv1, then using the result kernel to manipulate the original input will produce the same result, as can be proved by Formula(1).

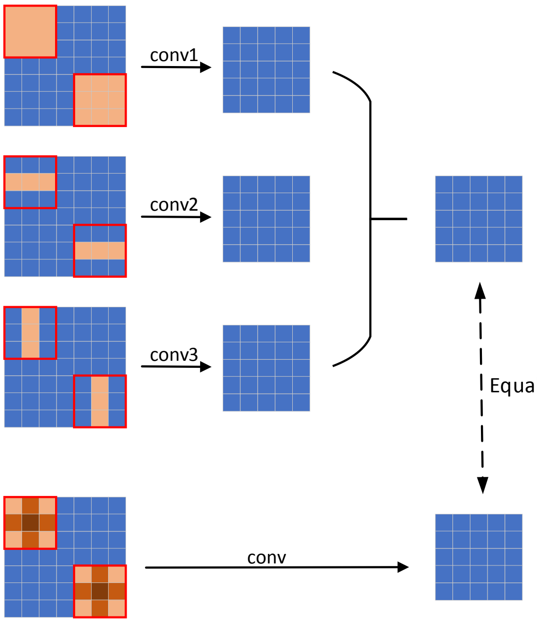

Figure 6.Additive property

During the training process, the ACB module splits the 3*3 convolution kernel into three different shapes of 3*3, 1*3 and 3*1 convolution kernels for training. But in the reasoning stage, it is merged into a 3*3 convolution, which does not bring the amount. However, the ACB module does not increase the amount of calculation relative to the 3*3 convolution, but compared with the original 1*1 convolution in the CSP module, it still increases the amount of parameters and calculation costs. Therefore, the CSPAC module is not used in the backbone feature extraction network to avoid excessive overhead, but only in the neck structure to replace the original CSP module, namely N4, N8 and N11. And the detection head of YOLOX is used for the detection of large objects, medium objects and small objects. The logo belongs to small object detection, and the logo in some angle images belongs to medium object. There is little demand for large object detection. H3 is a detection head for detecting large objects. The features processed by N14 are only used for H3 detection head. The features used by H1 and H2 have been processed before. In order to avoid too many parameters and calculations, N14 continues to use the previous CSP junction.

The CSPAC module is proposed to improve the neck structure of YOLOX, so that the network focuses on the learning of the convolution kernel skeleton part to better focus on local key features. Moreover, the CSPAC module is easier to converge with the network and reduces the cost of training the model.

3.3*Bicat*

The feature fusion network of YOLOX uses PANet. Compared with the FPN structure used by YOLOv3, it has a top-down feature fusion process. It is difficult to detect small objects in the target detection task. In the convolution process, the pixels of large objects are more than those of small objects. With the deepening of convolution, the features of large objects are easily retained, while the features of small objects are easily ignored with the deepening of network hierarchy. The FPN structure solves this problem to some extent. FPN takes the features of different resolutions generated in the previous step as input and outputs the fused features. In FPN, there is only a bottom-up feature fusion process. PANet establishes a new top-down path on the basis of FPN, which integrates shallow information and deep information more fully. Bidirectional Feature Pyramid Network (BiFPN)[28] deletes nodes with only one input edge. And add an additional edge between the input node and the output node in the same layer. In the Efficient Det network, BiFPN is reused many times to achieve a higher level of integration. If a multi-layer BiFPN structure is used in YOLOX, the network will be too complex. However, a prominent feature of BiFPN is that the feature fusion method used in the fusion process is weighted feature fusion. Therefore, the PANet structure is used to improve the neck network of YOLOX. Inspired by BiFPN, the original Concat feature fusion method is improved to weighted feature fusion, named Bicat.

Concat feature fusion directly splices the two features. If the dimensions of the two input features x and y are p and q, the dimension of the output feature z is p + q, and the feature fusion method treats features of different scales equally. The weighted feature fusion introduces weights, which can better balance the feature information of different scales. Different input features have different resolutions. The contribution of these inputs to the output features is usually unequal. Adding additional weights allows the network to learn the importance of each input feature. According to the characteristics of YOLOX feature fusion network, the designed Bicat feature fusion process is shown in Figure 7.

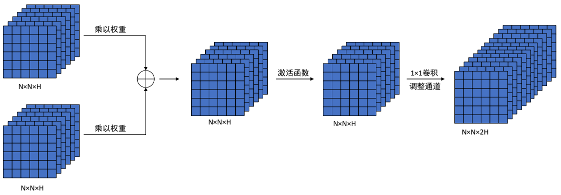

Figure 7.Bicat process

BiCat adopts Add fusion method, and the number of channels is halved compared with the previous Concat feature fusion method. In order to ensure that the subsequent feature channels are equal to the previous ones so that the subsequent network can process the features, it is necessary to perform feature mapping through 1*1 convolution while realizing channel expansion. The BiCat feature fusion calculation method is as Formula(2). Where wi is the learnable weight, and ε = 0.0001 is a small value to avoid numerical instability. I 1 and I 2 are channels of two feature layers.

$$F = \frac{w_1}{\varepsilon + \sum_i^2 w_i} \times I_1 + \frac{w_2}{\varepsilon + \sum_i^2 w_i} \times I_2 \qquad (2)$$

The Bicat feature fusion method can give higher weight to the features with stronger expression ability and improve the network pair. In addition, compared with the previous Concat feature fusion method, Bicat adds the weighted feature maps, and the amount of information describing each dimension of the image increases, but the dimension of the description image does not increase, so its calculation amount is much smaller than the Concat method.

IV. RESULTS

4.1*Experimental Environment and Parameter Description*

All models in this paper were implemented on the deep learning algorithm framework Pytorch 1.10.1. A GeForce RTX 4090 GPU was used for training and testing in all of our experiments. All ablation experiments were trained using the standard stochastic gradient descent method. The training process selects the SGD optimizer and uses the cos mode to update the learning rate. The initial learning rate was 0.01 and the weight attenuation was 0.0005. The batch size was set to 2, with 300epochs 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 (3).

$$\mathrm{AP} = \int_0^1 \mathrm{p(r)dr} \qquad (3)$$

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 (4).

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \qquad (4)$$

4.3*Dataset*

VLD-45[18], 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 45000 images and 50359 objects.

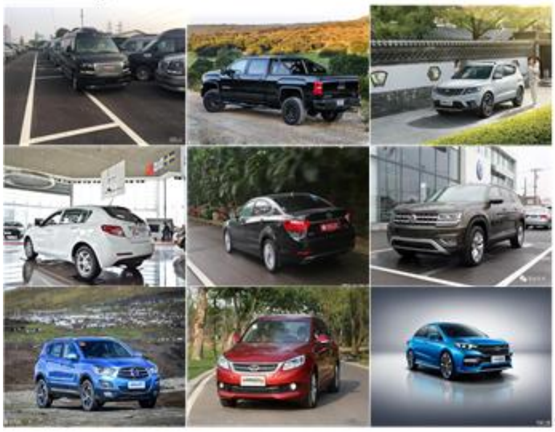

Figure 8.Examples from the dataset

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 30000 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 15000 images separately

4.4*Experimental Results and Analysis*

In order to evaluate the proposed improvement method, the experiment is carried out according to the following process. 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.

Firstly, in order to avoid the excessive addition of the channel attention mechanism, the detection effect is not ideal, and the ablation experiment is performed separately for the Bicat feature fusion method to determine its effectiveness. In the original YOLOX network, the Concat feature fusion method of the enhanced feature extraction part is improved to the Bicat method, and finally the Bicat method is applied to all positions that need feature fusion. Among them, N13 maintains the original Concat feature fusion method. Because the vehicle sign detection task belongs to small target detection, N12, N13, N14 and H3 do not play a big role in the detection task, so this part is no longer tested. The following 5 ablations were performed according to the above method.

1. YOLOX: The vehicle sign detection experiment is carried out on the original YOLOX without any improvement;

2. YOLOX-Bi-N3: The first up sampled feature in the original YOLOX is fused by Concat;

3. YOLOX-Bi-N7: The second up sampled feature in the original YOLOX is fused by Concat;

4.YOLOX-Bi-N10:The first down sampled feature in the original YOLOX is merged by Concat;

5. YOLOX-Bi: The feature fusion method in the original YOLOX is all by Concat except N13.

The above five improved models are tested on the same training set, verification set and test set. The results before and after improvement are shown in Table 1.

Table 1. Structure and result of ablation experimental model

Number	Experiment	Input	mAP(%)
1	YOLOX	416×416	58.35
2	YOLOX-Bi-N3	416×416	61.35
3	YOLOX-Bi-N7	416×416	61.30
4	YOLOX-Bi-N10	416×416	61.71
5	YOLOX-Bi	416×416	61.38

From the experimental results in Table 1 the detection effect of the model is improved after improving the N3, N7 and N10 in the original YOLOX network and all three parts into the Bicat feature fusion method. This is because Bicat imposes an attention mechanism when performing feature extraction, giving weight to the channel, so that the network has a focus during the training process. However, the effect of using BiCat in different feature fusion positions is different. In Experiment 2, the Concat of N3 is replaced by the Bicat feature fusion method. Compared with the original YOLOX detection results in Experiment 1, mAP increased by 3%, and the model detection effect was greatly improved. In Experiment 3, the Concat of N7 was replaced by Bicat, and mAP was increased by 2.95%. In Experiment 4, the Bicat feature fusion method was used to replace Concat at the N10 position, and the mAP was as high as 61.71%, an increase of 3.36%. In Experiment 5, Bicat feature fusion method was used at three positions of N3, N7 and N10, and a good improvement effect was also obtained. The mAP is increased by 3.03%, but its increase is not as high as that of Bicat only in N10. The improvement effect from high to low is: Bicat feature fusion is used only in N10, Bicat method was used in N3, N7 and N10, Bicat method is used only in N3, Bicat method is used only in N7. After analysis, there may be the following reasons. Firstly, the detection effect of Bicat feature fusion method in N3 is not much different from that in N7, while the detection method in N10 is significantly higher than that in the first two positions. At the N10 position, the previous up sampling and feature fusion operations have been performed, and the high-level semantic information and the low-level location information have been well fused. After that, simple attention is applied to the channel, which can make the network pay more attention to the effective channel, improve the network learning ability, and obtain the detection effect which is greatly improved compared with the previous one. Secondly, using the Bicat method at the three positions of N3, N7 and N10 is worse than using this method only at N10. It may be because that after the attention mechanism is applied at all locations, it is easy to cause over-fitting of the network, resulting in a decrease in the detection effect.

The above experiments and results show that the Bicat feature fusion method can greatly improve the detection results of vehicle sign detection tasks. And through the experimental data, it can be used for N10 position is the best way to improve, in the N10 position using Bicat feature fusion method can make the network better learning channel characteristics. When paying attention to deep features, we do not ignore the learning of shallow features. Shallow features are crucial to the detection of small targets. This method can improve the learning ability of small target features, thereby improving the detection accuracy of the model. In order to continue to verify the influence of other modules on the model detection effect, the following five ablation experiments were continued.

1. YOLOX;
2. YOLOX-Bi-N10;
3. YOLOX-ST: Replace C5 and C7 of the backbone feature extraction network with Swin Transformer Block;
4. YOLOX-AC: The proposed CSPAC module is introduced to replace the original CSP structure (N4, N8 and N11) in the network neck;
5. YOLOX-ST-AC-Bi(N3),the improved YOLOX model: Both Swin Transformer Block and CSPAC are introduced, and the Concat feature fusion method after the first down sampling of the neck network is replaced by the Bicat method.

The above five improved models were tested on the same training set, validation set and test set. The results before and after improvement are shown in Table 2.

Table2. Structure and result of ablation experimental model

Number	Experiment	Input	mAP(%)
1	YOLOX	416×416	58.35
2	YOLOX-ST	416×416	60.01
3	YOLOX-AC	416×416	61.98
4	YOLOX-Bi-N10	416×416	61.71
5	YOLOX-ST-AC-Bi	416×416	64.21

It can be seen from Table 4.2 that compared with the original YOLOX model, Experiment 2 replaced the C5 and C7 of the backbone feature extraction network with Swin Transformer Block, and the average accuracy was improved by 1.66%.The experimental results show that the application of Swin Transform Block can better extract global features. The improved backbone feature extraction network can combine global features and local features to reduce the influence of complex background on vehicle sign detection. In experiment 3, CSPAC module was introduced to replace the original CSP structure of N4, N8 and N11 in the network neck. This method can enrich spatial features and strengthen the extraction of vehicle sign information, and the average accuracy is increased by 2.63%.In experiment 4, after the bottom-up feature fusion of the neck of the YOLOX network, the original Concat feature fusion method is no longer used after the first down sampling, but the Bicat method is used instead. In the process of feature fusion, the effective features of the image are paid more attention by applying weight to the channel. This method brings a large increase in the detection results, and mAP is up to 3.36% higher than the original YOLOX. In Experiment 5, Swin Transformer Block, CSPAC and Bicat were introduced at the same time, and mAP was increased by 5.86%, which proved that the improved scheme was feasible.

Through the heat map, we can intuitively see the region of interest of the improved model to the image, as shown in Figure 9. It can be seen from the figure that the improved model can better focus on the characteristics of the logo position. Whether it is a vehicle image taken from the front or a vehicle image taken from the side or the rear, the model can pay attention to the characteristics of the vehicle logo. Therefore, the improved model can extract features in a targeted manner, achieve better detection results, and meet the needs of practical applications.

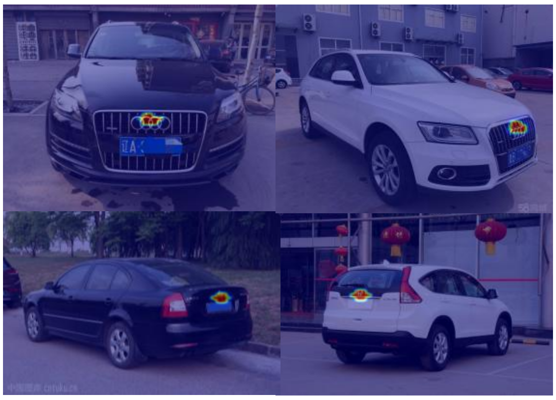

Figure 9.The heat map of the output image in YOLOX

In order to further verify the effectiveness of the improved model, the trained original YOLOX model and the improved model are used to detect the pictures in the test set on the VLD-45 data set selected in this paper. The network uses boxes to mark the size and location of the detected vehicle signs, and marks the classification and confidence of the target. The detection results are shown in Figure 10.

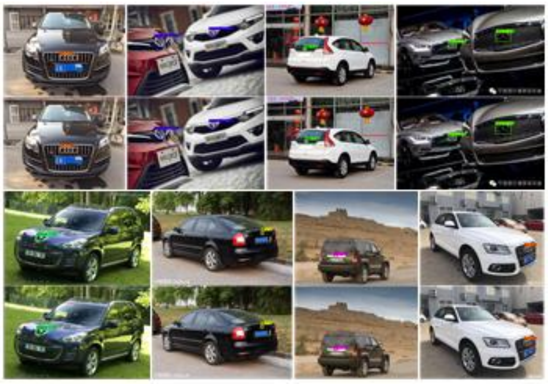

Figure 10. Comparison of detection effects before and after model improvement

In the figure, the first line and the third behavior are the original YOLOX detection results, and the second line and the fourth behavior are the improved model detection results. Although the original YOLOX basically does not appear error detection, missed detection, etc., the confidence of some targets is not high. The improved model inherits the good detection effect of the original YOLOX and improves the confidence of the detection.

In order to verify the detection effect of the model proposed in this paper, this study compares and tests the target detection model SSD[29], Retina Net[30], Free Anchor[31], Efficient Det, Center Net[32], YOLOv5[33]. Experiments and comparisons are also performed with TPH-YOLOv5[34], which is proposed to improve YOLOv5 for small target detection. The experimental results are shown in Table 2. The table lists the models that detect all categories of mAP in the VLD-45 dataset (IoU takes 0.5 to 0.95). Compared with other models, the improved YOLOX model has the highest average accuracy and the highest accuracy in vehicle sign detection tasks.

Table 2. Comparison of experimental model structure and results

Nubber	Experiment	Backbone	mAP(%)
1	SSD	VGG	49.87
2	RetinaNet	ResNet50	44.33
3	FreeAnchor	ResNet50	56.20
4	CenterNet	ResNet50	56.37
5	EfficientDet_d2	Effificient	59.00
6	YOLOv5	CSPDarknet	56.46
7	TPH-YOLOv5	CSPDarknet	62.20
8	MyModel	ST-Darknet	64.21

V. CONCLUSION

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. By introducing Swin Transformer Block, the backbone feature extraction network is promoted to better extract the global features of vehicle images. By introducing an asymmetric convolution structure into the neck structure of the network, the receptive field is expanded and the ability of the network to extract the key features of the vehicle logo is enhanced. Finally, based on BiFPN, a Bicat feature fusion method is designed to replace the Concat method. Through the model's learning of channel feature weights, the proportion weight of related channels is improved, and the number of detection model parameters is reduced. 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 5.86%.

REFERENCES

[1] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv: 2107.08430, 2021.

[2] Pan H, Zhang B. An integrative approach to accurate vehicle logo detection[J]. Journal of Electrical and Computer Engineering, 2013, 2013: 18-18.

[3] Llorca D F, Arroyo R, Sotelo M A. Vehicle logo recognition in traffic images using HOG features and SVM[C]//16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). IEEE, 2013: 2229-2234.

[4] Thubsaeng W, Kawewong A, Patanukhom K. Vehicle logo detection using convolutional neural network and pyramid of histogram of oriented gradients[C]//2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE, 2014: 34-39.

[5] Zhao J, Wang X. Vehicle-logo recognition based on modified HU invariant moments and SVM[J]. Multimedia Tools and Applications, 2019, 78: 75-97.

[6] Peng H, Wang X, Wang H, et al. Recognition of low-resolution logos in vehicle images based on statistical random sparse distribution[J]. IEEE transactions on intelligent transportation systems, 2014, 16(2): 681-691.

[7] Yu Y, Wang J, Lu J, et al. Vehicle logo recognition based on overlapping enhanced patterns of oriented edge magnitudes[J]. Computers & Electrical Engineering, 2018, 71: 273-283.

[8] Huang Y, Wu R, Sun Y, et al. Vehicle logo recognition system based on convolutional neural networks with a pretraining strategy[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 1951-1960.

[9] Li B, Hu X. Effective vehicle logo recognition in real-world application using mapreduce based convolutional neural networks with a pre-training strategy[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(3): 1985-1994.

[10] Yu Y, Guan H, Li D, et al. A cascaded deep convolutional network for vehicle logo recognition from frontal and rear images of vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 22(2): 758-771.

[11] Yu Y, Li H, Wang J, et al. A multilayer pyramid network based on learning for vehicle logo recognition[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(5): 3123-3134.

[12] Liu R, Han Q, Min W, et al. Vehicle logo recognition based on enhanced matching for small objects, constrained region and SSFPD network[J]. Sensors, 2019, 19(20): 4528.

[13] Lu W, Zhao H, He Q, et al. Category-consistent deep network learning for accurate vehicle logo recognition[J]. Neurocomputing, 2021, 463: 623-636.

[14] Xia Y, Feng J, Zhang B. Vehicle Logo Recognition and attributes prediction by multi-task learning with CNN[C]//2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2016: 668-672.

[15] Ke X, Du P. Vehicle logo recognition with small sample problem in complex scene based on data augmentation[J]. Mathematical Problems in Engineering, 2020, 2020.

[16] Liu J, Shen F, Wei M, et al. A Large-Scale Benchmark for Vehicle Logo Recognition[C]//2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC). IEEE, 2019: 479- 483.

[17] Zhang J, Yang S, Bo C, et al. Vehicle logo detection based on deep convolutional networks[J]. Computers & Electrical Engineering, 2021, 90: 107004.

[18] Yang S, Bo C, Zhang J, et al. VLD-45: A big dataset for vehicle logo recognition and detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(12): 25567-25573.

[19] Li W, Chen Y, Hu K, et al. Oriented reppoints for aerial object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1829-1838.

[20] Yang C, Huang Z, Wang N. Querydet: Cascaded sparse query for accelerating high-resolution small object detection[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022: 13668-13677.

[21] Akshatha K R, Karunakar A K, Shenoy S, et al. Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 195: 77-89.

[22] Zhong S, Zhou H, Ma Z, et al. Multiscale contrast enhancement method for small infrared target detection[J]. Optik, 2022, 271: 170134.

[23] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[24] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[25] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[26] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[27] Ding X, Guo Y, Ding G, et al. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1911-1920.

[28] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[29] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision– ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[30] CLin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[31] Zhang X, Wan F, Liu C, et al. Freeanchor: Learning to match anchors for visual object detection[J]. Advances in neural information processing systems, 2019, 32.

[32] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv:1904.07850, 2019.

[33] Jocher G, Stoken A, Borovec J, et al. ultralytics/yolov5: v5. 0-YOLOv5-P6 1280 models AWS Supervise. ly and YouTube integrations[J]. Zenodo, 2021, 11.

[34] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.