



Article

Vehicle Classification Algorithm Based on Improved Vision Transformer

Xinlong Dong ¹, Peicheng Shi ^{1,*}, Yueyue Tang ¹, Li Yang ¹, Aixi Yang ² and Taonian Liang ³

¹ School of Mechanical and Automotive Engineering, Anhui Polytechnic University, Wuhu 241000, China; dxl007ha@163.com (X.D.); 13053086906@163.com (Y.T.); yl1732460184@163.com (L.Y.)

² Polytechnic Institute, Zhejiang University, Hangzhou 310015, China; yangaixi@zju.edu.cn

³ Chery New Energy Automobile Co., Ltd., Wuhu 241000, China; ltn_99@126.com

* Correspondence: shipeicheng@ahpu.edu.cn

Abstract: Vehicle classification technology is one of the foundations in the field of automatic driving. With the development of deep learning technology, visual transformer structures based on attention mechanisms can represent global information quickly and effectively. However, due to direct image segmentation, local feature details and information will be lost. To solve this problem, we propose an improved vision transformer vehicle classification network (IND-ViT). Specifically, we first design a CNN-In D branch module to extract local features before image segmentation to make up for the loss of detail information in the vision transformer. Then, in order to solve the problem of misdetection caused by the large similarity of some vehicles, we propose a sparse attention module, which can screen out the discernible regions in the image and further improve the detailed feature representation ability of the model. Finally, this paper uses the contrast loss function to further increase the intra-class consistency and inter-class difference of classification features and improve the accuracy of vehicle classification recognition. Experimental results show that the accuracy of the proposed model on the datasets of vehicle classification BIT-Vehicles, CIFAR-10, Oxford Flower-102, and Caltech-101 is higher than that of the original vision transformer model. Respectively, it increased by 1.3%, 1.21%, 7.54%, and 3.60%; at the same time, it also met a certain real-time requirement to achieve a balance of accuracy and real time.



Citation: Dong, X.; Shi, P.; Tang, Y.; Yang, L.; Yang, A.; Liang, T. Vehicle Classification Algorithm Based on Improved Vision Transformer. *World Electr. Veh. J.* **2024**, *15*, 344. <https://doi.org/10.3390/wevj15080344>

Academic Editor: Peter Van den Bossche

Received: 8 July 2024
Revised: 17 July 2024
Accepted: 29 July 2024
Published: 30 July 2024



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

Keywords: vehicle classification; vision transformer; local detail features; sparse attention module; contrast loss

1. Introduction

Vehicle classification technology [1] is an important part of the driverless perception system, which is able to identify different types of vehicles, thereby helping vehicles make correct decisions and plan the road. Vehicle classification refers to the differentiation and identification of different types of vehicles, which can be divided into cars, trucks, motorcycles, etc. The development of vehicle classification technology has had a positive impact on the fields of transportation [2], automatic driving [3], target tracking [4], and artificial intelligence blockchain [5], and promoted the development of transportation and intelligent driving.

Early vehicle classification mainly used traditional machine learning algorithms, such as support vector machines [6], decision trees [7], etc., based on hand-designed feature representations, which were effective in some simple scenarios but had limitations in processing complex image data and long running time, which could not meet the real-time perception capabilities of autonomous driving. With the continuous development of modern convolutional neural networks [8] and artificial intelligence [9], vehicle classification technology has made great breakthroughs. The vehicle classification model based on deep learning [10,11] can automatically learn the high-level feature representation in the image, and the hierarchical structure makes it have excellent feature extraction ability, which

greatly improves the accuracy and robustness of vehicle classification. Through deep neural networks in deep learning, different features and tables can be learned from a large number of vehicle image data, which can identify and distinguish different types of vehicles, making vehicle classification more accurate and efficient. It has become the mainstream way of vehicle classification, providing the necessary perception capabilities for the field of autonomous driving. Although convolutional neural networks have excellent local feature capture performance in feature extraction of vehicle images, due to the scale sensitivity of convolutional operations, CNNs may not capture enough information and lack global feature representation ability, resulting in poor vehicle classification performance.

In recent years, the attention mechanism-based transformer structure [12] has been applied to the field of natural language processing (NLP) and has achieved great success and advantages in this field, which has inspired researchers to introduce the self-attention mechanism into computer vision tasks. The model uses the attention mechanism to focus attention on the region of interest when processing vehicle images, thereby improving the accuracy of vehicle classification. Dosovitskiy et al. [13] proposed a vision transformer (ViT) image classification model, proving that a transformer can be applied to the field of computer vision and has achieved success. ViT divides input images into small image blocks for processing, and each image block is regarded as an input sequence, passes through a series of transformer coding layers for feature representation learning, and then performs a multi-head self-attention operation on the feature sequence. The expressive ability and feature learning ability of the model are increased. Although the visual transformer has achieved good performance in computer vision tasks, it is also widely used for tasks such as image classification and object detection. However, directly segmenting images also brings some disadvantages and challenges, such as the ability to extract edge and local feature information is weakened, the image information is lost, and the context information is limited.

Although the accuracy of the vehicle classification network based on the vision transformer is improved compared with that of the convolutional neural network. However, due to the fact that the vision transformer processes the image by dividing the image into small pieces of fixed size, the sensitivity of the model to local details will be insufficient, some local feature information will be lost, and the detailed feature information in the vehicle image cannot be better extracted. At the same time, there are problems such as a large amount of computation and parameters, which will have a corresponding impact on the real-time nature of autonomous vehicles. In addition, a vision transformer may perform well on specific datasets, but its generalization ability may be limited in the face of new, unseen data distributions, which is especially important in practical applications such as autonomous driving. The hierarchical structure of a convolutional neural network can be more conducive to the complete extraction of local detail feature information from images. The most direct way to change this situation is to apply the convolutional neural network to the vision transformer model to capture the missing detail feature information of the vision transformer model so as to improve the accuracy of vehicle classification and enhance the environmental perception ability of autonomous vehicles.

Based on the above research, we propose an improved vision transformer vehicle classification model structure, IND-ViT, which uses a convolutional neural network to extract the detailed feature information of vehicle images, increase the feature expression ability of the model, and improve the accuracy of vehicle classification. The architecture proposed in this paper has the following contributions:

1. We propose an improved vision transformer vehicle classification network, IND-ViT, which designs a local information feature extraction module to make up for the local detail features lost by the vision transformer's direct image segmentation operation and improve the model's perception ability.
2. Aiming at the problem of misdetection caused by the large similarity of some vehicles, we proposed a sparse attention module based on the attention mechanism, which comprehensively utilized the attention weight information of all coding layers to

- capture the discernable region in the image and further improved the fine-grained feature representation ability of the model.
3. We refer to the contrast loss function to further increase the intra-class consistency and inter-class difference of network learning features. Contrast loss makes the similarity of classification features corresponding to different labels minimum and the similarity of classification features corresponding to the same labels maximum.
 4. Through extensive testing experiments on datasets such as BIT-Vehicles, CIFAR-10, Oxford Flower-102, and Caltech-101, the results show that, compared with the original ViT network, the accuracy of the improved method in this paper is increased by 1.3%, 1.21%, 7.54%, and 3.60%, respectively, which is superior to other mainstream methods.

2. Related Works

2.1. Vehicle Classification Based on Convolutional Neural Network

With the rapid development of artificial intelligence, convolutional neural networks (CNNs) have always been considered the basic models of computer vision, providing powerful feature extraction and learning capabilities for vehicle classification technology, and CNNs can automatically learn representations about objects from vehicle images. Researchers have developed many advanced vehicle classification models by combining CNNs with vehicle classification technology, which has advanced the development of the field of intelligent transportation. Maungmai et al. [14] proposed a convolutional neural network-based vehicle classifier to improve the accuracy of vehicle classification by classifying the color detail information of vehicle images. Yu et al. [15] proposed a deep learning model that combines fast R-CNN vehicle detection, CNN feature extraction, and joint Bayesian network classification to achieve fine-grained vehicle classification in complex traffic scenarios. Ma et al. [16] proposed a channel maximum pooling (CMP) scheme to optimize feature extraction by inserting a new layer between the fully connected layer and the convolutional layer, reduce the number of parameters, and improve the classification accuracy and CNN generalization ability in fine-grained vehicle classification tasks. Jo et al. [17] proposed a transfer learning-based vehicle classification method to achieve effective classification of vehicle models on limited-scale datasets. Neupane et al. [18] proposed a simplified CNN-based model for classifying vehicles in low-resolution surveillance images, with the potential to achieve high accuracy with low complexity in low-quality images, providing a new idea for the application of standard low-cost cameras to enhance the application of intelligent transportation systems. Hasanvand et al. [19] proposed an SVM vehicle classification algorithm, which uses the unified collected car images for corresponding noise processing and provides four classifiers (support vector machine, k-nearest neighbor, perceptron neural network, and Bayesian decision theory) for vehicle classification through background removal and feature extraction so as to improve the accuracy of vehicle type recognition and effectively assist traffic violation identification and management. At present, the road condition monitoring technology of camera sensors provides favorable information for vehicle classification and control systems. To this end, Zhao et al. [20] created a large-scale road surface image dataset and trained a CNN classification model, enhanced the robustness of the algorithm through the improved Dempster-Shafer evidence theory, and deployed the developed model on the embedded hardware platform. Although convolutional neural networks have excellent local feature capture performance in feature extraction, due to the scale sensitivity of convolution operations, CNNs may not capture enough information and lack global feature representation capabilities, resulting in poor vehicle classification performance.

2.2. Vehicle Classification Based on Attention Mechanism

The vehicle classification model based on the attention mechanism [12] can effectively capture the contextual information in the image, which is helpful to improve the understanding of the relationship between the target and its surroundings in the intelligent transportation system. Models can focus attention on areas of interest when processing

vehicle images, improving the accuracy of vehicle classification. Zhao et al. [21] proposed a vehicle classification model combined with a visual attention mechanism, which uses the visual attention module to enhance the key parts of the vehicle image and suppress the unimportant parts to form a focused image, thereby improving the accuracy of vehicle classification. PVT [22] designed a progressive attenuation pyramid structure and a spatial reduction attention mechanism, which reduced the length of the input sequence through the hierarchical attenuation of image resolution and reduced the computational cost of the model. The attention mechanism-based vision transformer classification model [13] has demonstrated excellent performance in a variety of image classification tasks, and its advantage lies in its ability to handle long-distance dependencies and capture complex interactions between different regions in the image. However, directly segmenting the image also brings some disadvantages and challenges, and the ability to extract edge and local feature information is weakened. To solve this problem, Zhu et al. [23] used a more complex image chunking strategy to deform convolution to improve the capture of local information of the image, but at the expense of computation. Chen et al. [24] proposed the Visformer model, which is a friendly transformer architecture with convolutional operators that achieves high recognition performance through step-by-step operation and stage design but is difficult to cope with object classification and recognition in complex environments. The ConViT model recently proposed by Stephane d'Ascoli et al. [25] has achieved good improvement results by improving the self-attention layer to enable the ViT model to obtain global information more fully. Liu et al. [26] proposed a Swin transformer model that uses sliding windows to model global information, which reduces the sequence length and increases efficiency but lacks the ability to extract local information. The conformer [27] model designs parallel CNN and transformer branches and uses a bridging module to achieve feature fusion, which will lead to the degradation of the image resolution of the upstream task due to the lack of position encoding of the image. The Levit model proposed by Graham et al. [28] uses cascading multiple small convolutions to obtain the local features of the image before the image is partitioned and at the same time increases the convolution step size for downsampling, which effectively reduces the number of parameters of the model, but the classification accuracy is not very high. Due to the limitations of visual transformer technology in multi-scale feature extraction and understanding of undisciplined traffic environments, Deshmukh et al. [29] proposed a vehicle detection framework STVD based on Swin transformers in irregular traffic environments, which uses Swin transformers and bidirectional feature pyramid networks (BIFPNs) to enhance multi-scale feature extraction capabilities. Using the fully connected vehicle detection head (FCVDH) to optimize vehicle size matching, STVD achieves high detection accuracy on multiple real-world traffic datasets. There are also researchers who include ViT in the HSI classification. Roy et al. [30] proposed a novel transformer architecture called morph-former, which enhances the feature interaction in hyperspectral image (HSI) classification by combining spectral and spatial morphological convolution operations and self-attention mechanisms, providing corresponding theoretical support for the field of vehicle classification. Based on the above research, we propose an improved vision transformer model for vehicle classification, which is different from the above literature.

3. Improved Vehicle Classification Model of Vision Transformer (IND-ViT)

Figure 1 shows the vehicle classification and recognition network framework based on the improved vision transformer proposed in this paper. Firstly, the input vehicle images are double-branched; one branch uses the convolutional neural network CNN-In D structure to extract local features to make up for the loss of detail information in the image block operation, and the other branch divides the image into N image blocks of $p \times p$ size, and then the image blocks are linearly mapped into serialized embedding vectors, and learnable classification vectors and position coding information are added. Secondly, the embedding vector is input into multiple stacked coding modules for feature extraction, and before the last layer of coding modules, the sparse attention module is used to find

the distinguishing pixel blocks, the discognizable regions, and their corresponding hidden features in the image. Then, the information extracted by the encoder and the local detail features extracted by CNN-In D were fused with feature information, and a global average pooling layer (GAP) was input for dimensionality reduction, and then the Softmax function was used to identify and classify the image. Finally, the class information of the vehicle is obtained, and the improved model is shown in Figure 1.

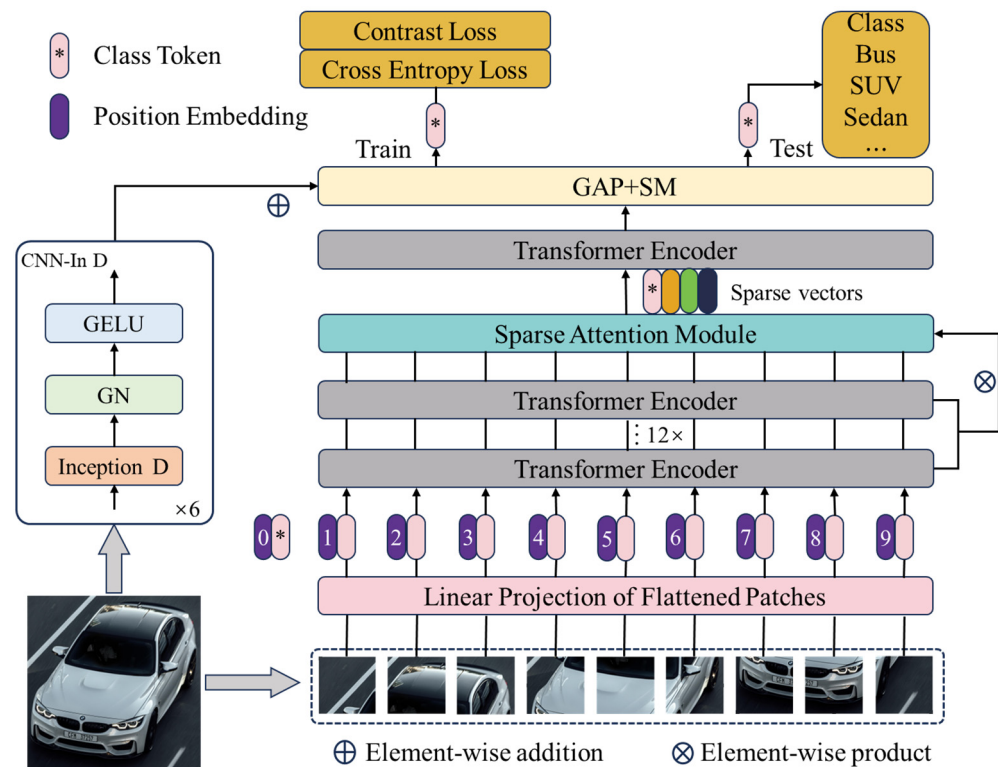


Figure 1. Vehicle classification and recognition network framework based on improved vision transformer.

3.1. Local Detail Feature Extraction Module: Inception D

The local detail feature extraction branch we designed consists of 6 CNN-In D modules, each of which consists of an Inception D module, a normalization layer (GN), and a nonlinear activation function (GELU). When designing the local detail feature extraction module, this paper draws on the ideas of Inception V1 and Inception V3 [31], uses convolution checks of different sizes to extract features from input feature maps, and then splices all operation results to make the network more adaptable and the feature information of each layer more abundant. Since the number of layers of the recognition model designed is relatively small, the Inception D module designed in this paper first uses 1×1 convolution check channels to compress the number of parameters of the model and then further reduces the number of parameters of the model through asymmetric convolution. The Inception D module used in the construction of the convolutional neural network recognition model in this paper is shown in Figure 2.

It can be seen from Figure 2 that the Inception D module consists of four parts. The first part is the convolution layer; the convolution kernel size is 1×1 . By using the convolution kernel size of 1×1 , the number of channels can be reduced or increased without changing the image size, so as to reduce the computational complexity of the model. The second part is the asymmetric convolution kernel 1×3 , 3×1 , which can capture local features and provide a large receptive field. The third part is the asymmetric convolution kernel 1×5 , 5×1 , which further expands the receptive field and captures a larger range of features. The middle two lines will do 1×1 convolution on the input first to reduce the number of input channels and reduce the complexity of the model. The fourth part starts with the average pooling layer; the size of the pooling kernel is 3×3 , and then the convolution layer; the

size of the convolution kernel is 1×1 . By using the maximum pooling operation of 3×3 , the image size can be reduced in the spatial dimension, and the output of the above four parts can be spliced in the channel dimension. The final output of the Inception D module is formed and fed into the next layer. By stacking multiple Inception modules in the model, you can build a deeper network architecture and improve the expressiveness of the model. We have demonstrated through numerous experiments that the best results are achieved when six Inception D modules are used, which not only increases the depth of the network but also reduces the risk of overfitting the model. In this way, the convolutional neural network can combine various feature information of low-level features and high-level features so that the high-level feature map contains both detailed information such as location and high-level information such as semantics, which enhances the ability of the network to express features and improves the recognition accuracy of the network.

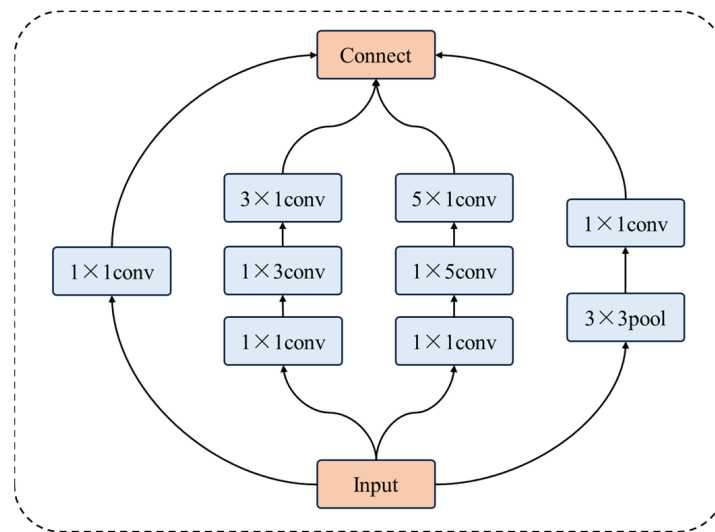


Figure 2. Extraction of local feature details: Inception D module.

3.2. Image Partitioning and Location Coding

The input received by the vision transformer model is serialized data, so the image needs to be divided into image blocks and mapped linearly to serialized vectors. When the input image size is $H \times W \times C$ and the image block size is p , the vision transformer model will divide the image into $p \times p$ non-overlapping N pixel blocks, as shown in Figure 3.

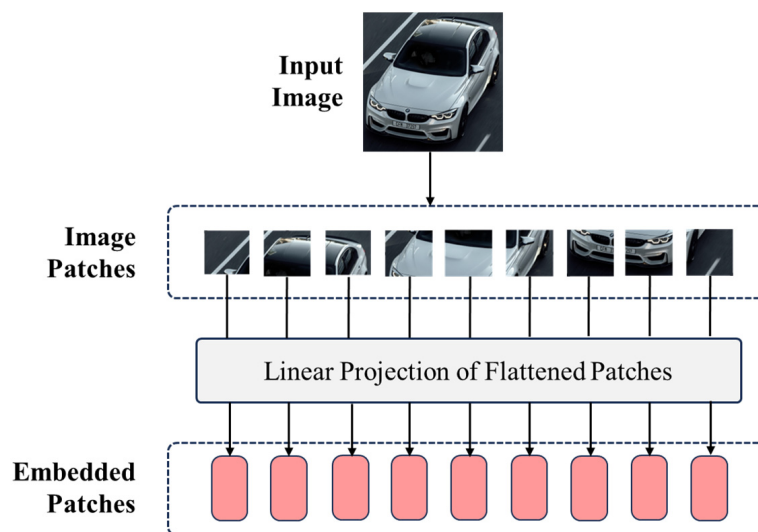


Figure 3. Image block processing.

After the image partition is completed, it is necessary to convert the 2D image block into a 1D sequence vector, first flattening the image block into a set of vectors, and then mapping it to the dimension size of D by linear transformation. Since the embedded vector does not contain position information, a special learnable position code needs to be added. In addition, the learnable classification vector is added as the final output feature for image classification. The embedded sequence data z_0 is shown in Equation (1), where E is the projection matrix, E_{pos} is the position encoding, and x_{class} is the classification vector.

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos}, E \in \mathbb{R}^{p^2 C \times D}, E_{pos} \in \mathbb{R}^{(N+1) \times D} \quad (1)$$

3.3. Encoder

The encoder of the vision transformer is stacked with L coding modules with the same structure. The coding module structure is shown in Figure 4. The coding module includes multi-head self-attention (MSA) and multi-layer perceptron (MLP). The multi-head self-attention module consists of N_h single-head self-attention (SA) units. For a single-headed self-attention unit, the first input is $z_p \in \mathbb{R}^{(N+1) \times D}$, and the input is transformed linearly to obtain query matrix Q , key matrix K , and value matrix V . The linear transformation is shown in Equations (2)–(4):

$$Q = z_p W^Q, W^Q \in \mathbb{R}^D \times d_K \quad (2)$$

$$K = z_p W^K, W^K \in \mathbb{R}^D \times d_K \quad (3)$$

$$V = z_p W^V, W^V \in \mathbb{R}^D \times d_K \quad (4)$$

where $d_K = \frac{D}{N_h}$, after Q , K , and V are obtained, attention weight matrix A is calculated as follows:

$$A = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_K}}\right), A \in \mathbb{R}^{(N+1) \times (N+1)} \quad (5)$$

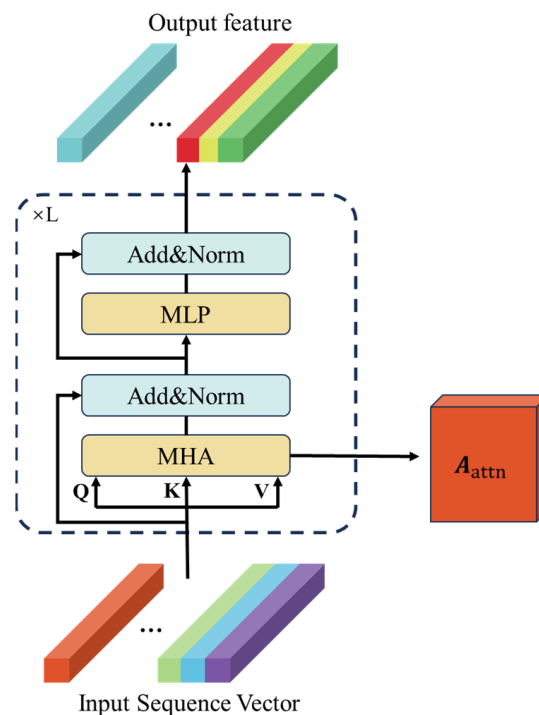


Figure 4. Vision transformer encoder module structure.

The element A_{ij} in matrix \mathbf{A} represents the correlation between the i th feature and the j th feature, the larger the value, the stronger the correlation, and $\sqrt{d_K}$ is the scaling factor. The output \mathbf{z}' of the single-head self-attention unit is obtained by multiplying the attention weight \mathbf{A} matrix by the value matrix \mathbf{V} :

$$\mathbf{z}' = \mathbf{A} \cdot \mathbf{V}, \mathbf{z}' \in \mathbb{R}^{(N+1) \times d_K} \quad (6)$$

Different single-headed self-attention units learn the relevant features in the independent feature subspace, and finally the multi-headed self-attention module splices the output results of the single-headed self-attention unit and then gets the output of the module through linear transformation. This output is resistorally linked with \mathbf{z}_p and used as input to the next multilayer perceptron module after layer normalization (LN).

$$MSA(\mathbf{z}_p) = \text{concat}_{i \in N_h} \left(SA(\mathbf{z}_p^i) \right) \mathbf{W}_{out} + \mathbf{b}_{out} \quad (7)$$

where $\mathbf{W}_{out} \in \mathbb{R}^D \times D$ is the weight and $\mathbf{b}_{out} \in \mathbb{R}^{(N+1) \times D}$ is the bias.

The multi-layer perceptron module uses two fully connected layers; the first fully connected layer uses the ReLU activation function, and the second does not use the activation function. The calculation formula is as follows:

$$MLP(\mathbf{X}) = \text{ReLU}(\mathbf{X} \cdot \mathbf{W}_1 + \mathbf{b}_1) \cdot \mathbf{W}_2 + \mathbf{b}_2 \quad (8)$$

If \mathbf{z}_{p-1} is the input of the p -th coding module, the output of the coding module is as follows:

$$\mathbf{z}'_p = \text{LN}(MSA(\mathbf{z}_{p-1}) + \mathbf{z}_{p-1}) \quad (9)$$

$$\mathbf{z}_p = \text{LN}(MLP(\mathbf{z}'_p) + \mathbf{z}'_p) \quad (10)$$

3.4. Sparse Attention Module

The key problem of vehicle image classification is whether the identification area in the image can be accurately located. We selected several vehicle model photos from the BIT-Vehicles dataset as examples to illustrate the nuances and recognizable areas between them. Take the sedan and SUV in Figure 5 as an example. The differences between the two types of vehicles are relatively subtle, mainly concentrated in the front part, while the difference between the SUV and the microbus is mainly the length of the body. In convolutional neural networks, discernable regions in images are mainly located through regional recommendation networks or weakly supervised segmentation masks, while in the vision transformer model, the multi-head self-attention mechanism can independently learn the weights of different image blocks. In order to make full use of this weight information to locate the discernable region, a sparse attention module (SAM) is proposed in this paper, as shown in Figure 6.

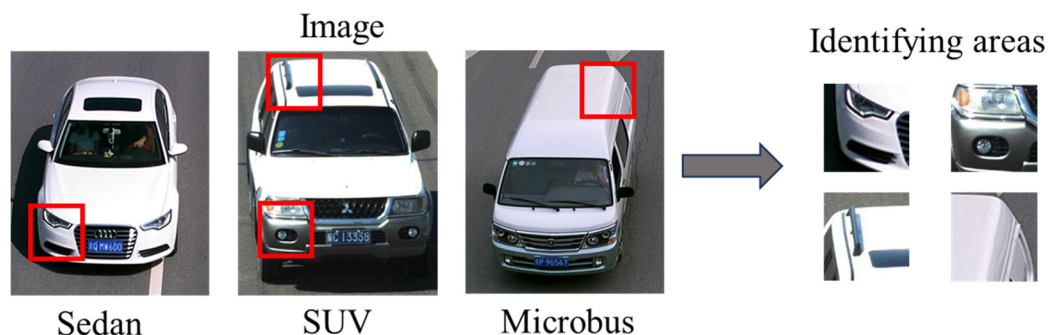


Figure 5. Identification regions of different categories of vehicles.

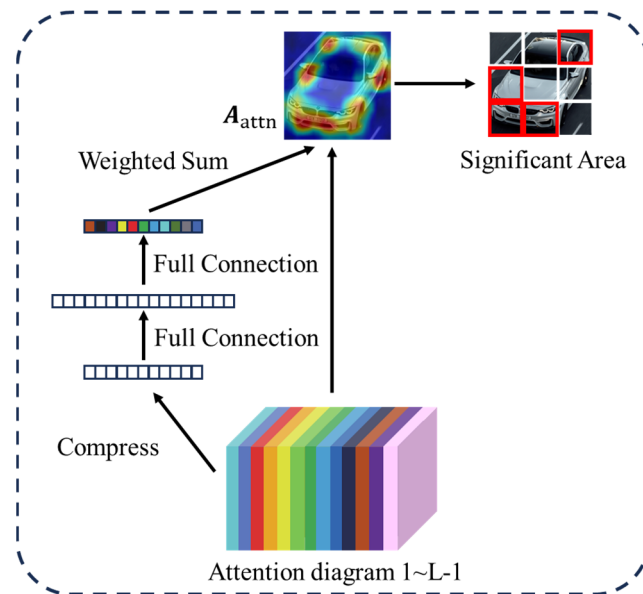


Figure 6. Sparse attention module.

If the vision transformer network contains L coding modules, the sparse attention module uses the weights learned at the first $L - 1$ coding layer to screen the hidden features $z_{L-1} = [z_{L-1}^1; z_{L-1}^2; \dots; z_{L-1}^{N_h}]$ input at the last coding layer. The weights learned by the first $L - 1$ coding layer are shown in Equations (11) and (12).

$$A_l = [A_l^1, A_l^2, \dots, A_l^{N_h}], l \in 1, 2, \dots, L - 1 \tag{11}$$

$$A_l^i = [a_l^{class}; a_l^1; a_l^2; \dots; a_l^N], l \in 1, 2, \dots, N_h, a_l^i \in \mathbb{R}^1 \times (N+1) \tag{12}$$

Due to the abstraction of high-level features, the attention map may not necessarily represent the importance of the corresponding input image block, so we use the attention map information of all previous coding modules and combine the compression excitation module to independently learn the weight of each attention map. The module first averages the attention map into a descriptor, then models the correlation between the attention maps using two fully connected layers, and finally obtains the weight value α for each attention map. After the weight value is normalized and weighted with the attention diagram, the final attention weight A_{attn} is obtained, as shown in Equation (13). The entire process is shown in Figure 6.

$$A_{attn} = \sum_{i=1}^{L-1} \alpha_i A_i \tag{13}$$

A_{attn} contains all the attention weight information of lower-layer features and higher-layer features and is more suitable for screening discernable regions than single-layer attention weight A_{L-1} . We use the weight $A_{attn}^{class} = [a_{final}^1; a_{final}^2; \dots; a_{final}^{N_h}]$ corresponding to the classification vector in A_{attn} to screen out the hidden feature corresponding to the largest weight among N_h self-attention heads. Finally, these hidden features are combined with the classification vector as the input of the last layer coding module.

$$z_{L-1}^{attn} = [z_{L-1}^{class}; z_{L-1}^{\alpha_1}; z_{L-1}^{\alpha_2}; \dots; z_{L-1}^{\alpha_{N_h}}] \tag{14}$$

The sparse attention module replaces all sequence vectors with feature vectors corresponding to the identification region, splicing them with classification vectors, and then inputs them into the last coding module, which not only retains the global classification feature information but also forces the last coding layer to pay attention to the subtle

differences between different categories. At the same time, give up a lot of low degree of differentiation of regional information, such as background and the superclass common characteristics, so as to improve the expression ability of detailed features of the network.

3.5. Loss Function

Like a vision transformer, we use the first vector of the network output, the component vector, for image classification. The loss function of the network includes cross-entropy loss L_{cross} and contrast loss L_{con} , as shown in Equation (15):

$$L = L_{cross}(\mathbf{y}, \mathbf{y}') + L_{con}(\mathbf{z}) \quad (15)$$

Cross-entropy loss is used to measure the similarity between the real label \mathbf{y} and the network prediction label \mathbf{y}' , defined as Equation (16):

$$L_{cross}(\mathbf{y}, \mathbf{y}') = -\sum_{i=1}^C y_i \log(y'_i) \quad (16)$$

In order to further increase the intra-class similarity and inter-class difference of network extraction features, we add contrast loss L_{con} . The comparison loss makes the similarity of classification features corresponding to different labels minimum, and the similarity of classification features corresponding to the same labels maximum. In order to balance positive and negative samples and prevent loss from being dominated by simple negative samples (different class features with small similarity), we introduce the threshold t_{con} , and only when the similarity of sample features of different classes is greater than t_{con} can it be included in the loss. When the batch size of input data is N , the comparison loss is defined as shown in Equation (17):

$$L_{con} = \frac{1}{N^2} \left[\sum_{j:y_i=y_j} \left(1 - \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|} \right) + \sum_{j:y_i \neq y_j} \max \left(\frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|} - t_{con}, 0 \right) \right] \quad (17)$$

4. Experimental Results and Analysis

4.1. Datasets and Evaluation Indicators

In order to verify the effectiveness of the IND-ViT network, we conducted an experimental evaluation based on the BIT-Vehicles dataset, which is a screenshot of surveillance video from real traffic produced by Beijing Institute of Technology. It includes six types of vehicles, including trucks, cars, SUVs, minivans, buses, and minibuses, with a total of 9850 images. At the same time, experiments are also carried out on public data sets CIFAR-10, Oxford Flower-102, and Caltech-101 to further verify the rationality of the network. The CIFAR-10 dataset includes 10 different types of images: planes, cars, birds, cats, deer, dogs, frogs, horses, boats, and trucks, and the dataset contains 60,000 images with a color image size of 32×32 . Oxford Flower-102 is an image dataset of 102 flower categories common in the UK, each containing 40 to 258 images, each with large scale, pose, and light variations, for a total of 8189 images. Caltech-101 consists of 101 object pictures, each category labeled with a single object, and each category contains approximately 40 to 800 images, varying in size, for a total of 8677 images. The above datasets were randomly divided into a training set, verification set, and test set according to the ratio of 6:2:2. The data expansion strategy in this experiment uses only random clipping and random horizontal flipping.

In order to quantitatively evaluate the performance of the classification algorithm, we use such evaluation indicators as precision, recall, and accuracy, which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (20)$$

where TP is the number of positive samples correctly predicted as positive class; FP is the number of negative samples incorrectly predicted as positive; TN is the number of negative samples correctly predicted as a negative class; FN is the number of positive samples incorrectly predicted to be negative classes.

4.2. Experimental Environment

The training and reasoning of this method were carried out on a Ubuntu18.04 server equipped with I7-10700 CPU and GeForce RTX 3060 GPU. The development language was Python3.8, and the relevant code was developed and written using the PyTorch deep learning framework. In the experiment, the image size of the public data set was 32×32 , the number of images for each input model training was set to 32, and the number of iterations was 150 rounds.

4.3. Training Parameters

The optimizer trained by the experimental model in this paper is AdamW [32], the learning rate is 0.0001, the weight decay rate is 0.01, and the number of iterations is 100. In order to accelerate the convergence speed of the model, this paper adopts the learning rate cosine decay period to dynamically adjust the learning rate and sets 20 iterations as one cycle.

4.4. Comparative Experiment

In order to verify the effect of the improved IND-ViT model in this paper, we not only tested the improved IND-ViT model but also tested the traditional CNN network model, the original transformer model, and the transformer fusion convolution model. ResNet-50 [33], ViT [13], Visex-Tiny [24], Levit-256 [28], ResT-lite [34], and the improved IND-ViT model are respectively adopted as classification network models of BIT-Vehicles data sets. The performance of other methods is experimentally obtained by copying their official code. The bold numbers in Table 1 represent the data for the best-performing method in that row.

Table 1. Performance comparison for vehicle classification on the BIT-Vehicles test set.

Accuracy	ResNet-50 [33]	ViT [13]	Visformer-Tiny [24]	Levit-256 [28]	CSWin-T [34]	IND-ViT
Track	96.30%	95.80%	96.20%	96.00%	97.00%	97.60%
Sedan	98.50%	98.40%	98.50%	98.90%	98.20%	99.20%
SUV	84.60%	84.30%	84.70%	85.60%	85.80%	87.20%
Minivan	99.80%	99.90%	99.95%	99.85%	99.90%	99.99%
Bus	99.80%	99.90%	99.90%	99.99%	99.95%	99.99%
Microbus	96.60%	96.20%	97.00%	96.40%	96.80%	98.30%
Average accuracy	95.93%	95.75%	96.05%	96.12%	96.28%	97.05%

As can be seen from Table 1, the traditional CNN network is difficult to carry out global representation capabilities, and the transformer model will ignore local feature details. As a result, the attention mechanism-based transformer model has a small difference in classification accuracy of various vehicles compared with the traditional CNN model. Although other methods based on transformer architecture fusion convolution can better improve the classification accuracy of ViT, the classification accuracy is not good compared with the traditional CNN network. Compared with the original ViT model, the accuracy of the improved model in individual vehicle classification and overall classification has been significantly improved, among which track has been improved by 1.8%, sedan by 1.2%, SUV by 2.9%, microbus by 2.1%, and the average accuracy has been improved by 1.3%. We also test the recognition time of a single image by testing the network to determine

whether the network can meet the real-time requirements, and the recognition time of a single image under different network models is shown in Table 2.

Table 2. Single image recognition time.

Network Model	Time/ms
ResNet-50 [33]	8.6372
ViT [13]	9.5764
Visformer-tiny [24]	10.6314
Levit-256 [28]	12.0295
CSWin-T [34]	11.1245
IND-ViT	10.8236

As can be seen from Table 2, the recognition time of the ResNet model based on CNN feature extraction is the shortest and fastest, followed by the ViT model based on transformer feature extraction. Compared with other improved ViT networks, the recognition speed of the proposed method is higher and can meet the real-time requirements. This paper uses confusion matrix to evaluate the model. The confusion matrix of the original ViT network model is shown in Figure 7.

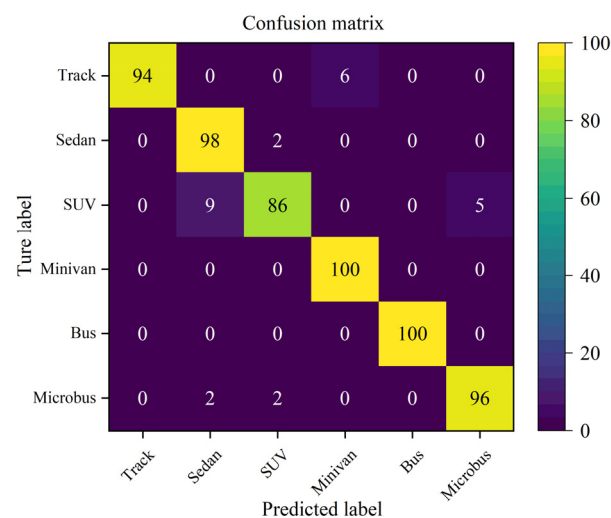


Figure 7. Identification confusion matrix of ViT network model.

It can be clearly seen from Figure 7 that the prediction accuracy of ViT in the original network is the highest for minivan and bus, followed by sedan, microbus, and track. In the original network, the recall rate of SUV is the lowest, and the main error is divided into sedan and microbus; the missed detection rate is 9% and 5%, respectively, followed by the wrong track being divided into minivan, the missed detection rate is 6%. Due to the similar appearance of SUVs and cars, it is easy for the network to confuse the two in classification, and the output prediction score is also very similar, which is easy to produce accidental errors. However, the lack of local feature details in the ViT model will lead to wrong classification. Therefore, the improved IND-ViT network in this paper adds a local feature information extraction layer and introduces a sparse attention module so that the network can obtain sufficient image details and discernible regions. The confusion matrix of the improved IND-ViT network is shown in Figure 8.

As can be seen from Figure 8, compared with the original network, the improved network using the proposed method has a 2% increase in the classification accuracy of microbus, a 4% decrease in the missed detection rate of truck classified as minivan, and a 4% increase in the classification accuracy rate of SUV, among which the missed detection rate of SUV classified as minivan and a 3% decrease in the detection rate of sedan. It shows that the improved network has a better classification effect in vehicle classification.

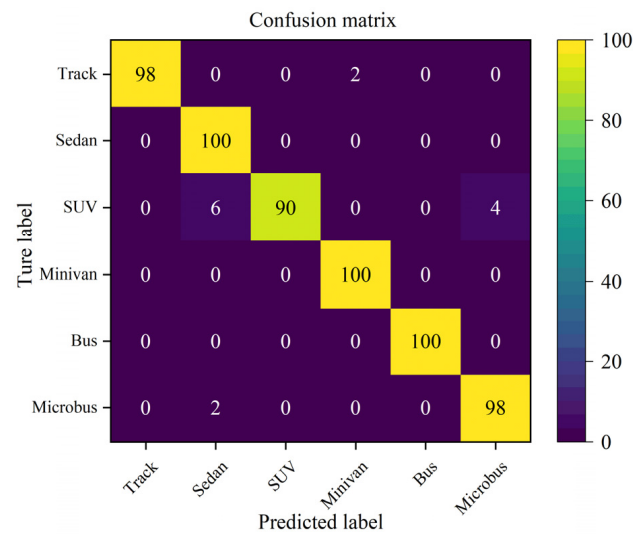


Figure 8. Identification confusion matrix of IND-ViT network model.

In order to further verify the effectiveness of the network proposed in this paper, we also conducted experiments on CIFAR-10, Oxford Flower-102, and Caltech-101. It is not difficult to find from Table 3 that the tests on the data set CIFAR-10, The classification accuracy of the proposed IND-ViT model is 1.21% higher than that of the original ViT model, reaching the highest 99.63%. For the Oxford Flower-102 data set, although the transformer model fusion convolution method in other literature can improve the accuracy of ViT, the classification accuracy still cannot reach the result of the traditional CNN model. The classification accuracy of the IIN-ViT model proposed in this paper is 7.54% higher than that of the original ViT model, 4.62% higher than ResNet-18, 1.68% higher than ResNet-50, and has better advantages than the traditional CNN structure. On the Caltech-101 dataset, the classification accuracy of the IND-ViT model is about 8% higher than that of the traditional CNN model, which is significantly better than other transformer architecture models. This paper further proves the accuracy of our proposed improved vision transformer vehicle classification model, IND-ViT, which makes full use of local detail features and high-level global features to enrich the network model with rich feature information and improve the classification accuracy of the model.

Table 3. Comparison of model performance on the CIFAR-10, Oxford Flower-102, and Caltech-101 datasets.

Models	CIFAR-10	OxfordFlowers-102	Caltech-101
ResNet-18 [33]	0.9328	0.7559	0.6401
ResNet-50 [33]	0.9435	0.7953	0.6419
ViT [13]	0.9842	0.7267	0.6859
Visformer-tiny [24]	0.9644	0.7216	0.5304
Levit-256 [28]	0.9651	0.7185	0.4619
CSWin-T [34]	0.9788	0.7678	0.5950
IND-ViT	0.9963	0.8021	0.7219

4.5. Qualitative Analysis

In order to qualitatively evaluate the recognition performance of the IND-ViT model and the effectiveness of each module, we used Grad-CAM [35] to calculate the attention heat maps of different models on images with multiple flowers for IND-ViT, ResNet-18, and ViT and carried out visual analysis of the heat maps, the results of which were shown in Figure 9. According to Figure 9, compared with ViT and ResNet-18, the IND-ViT model can identify the discernible region of the vehicle more accurately, which can prove the validity of our sparse attention module. In addition, ResNet-18 can accurately identify the

flower in the center of the image but not other flowers. While the ViT model can feel where all the flowers are, it cannot get more precise details. Compared with the results identified by the ResNet-18 and ViT models, the IND-ViT model can not only sense the location of all flowers but also obtain key local details of flowers. Therefore, IND-ViT can obtain a global representation of features through the self-attention module of the transformer branch, as well as local details from the convolutional module of the CNN branch, thus effectively merging local information and global information. This further proves the effectiveness of our local detail feature extraction module and sparse attention module.

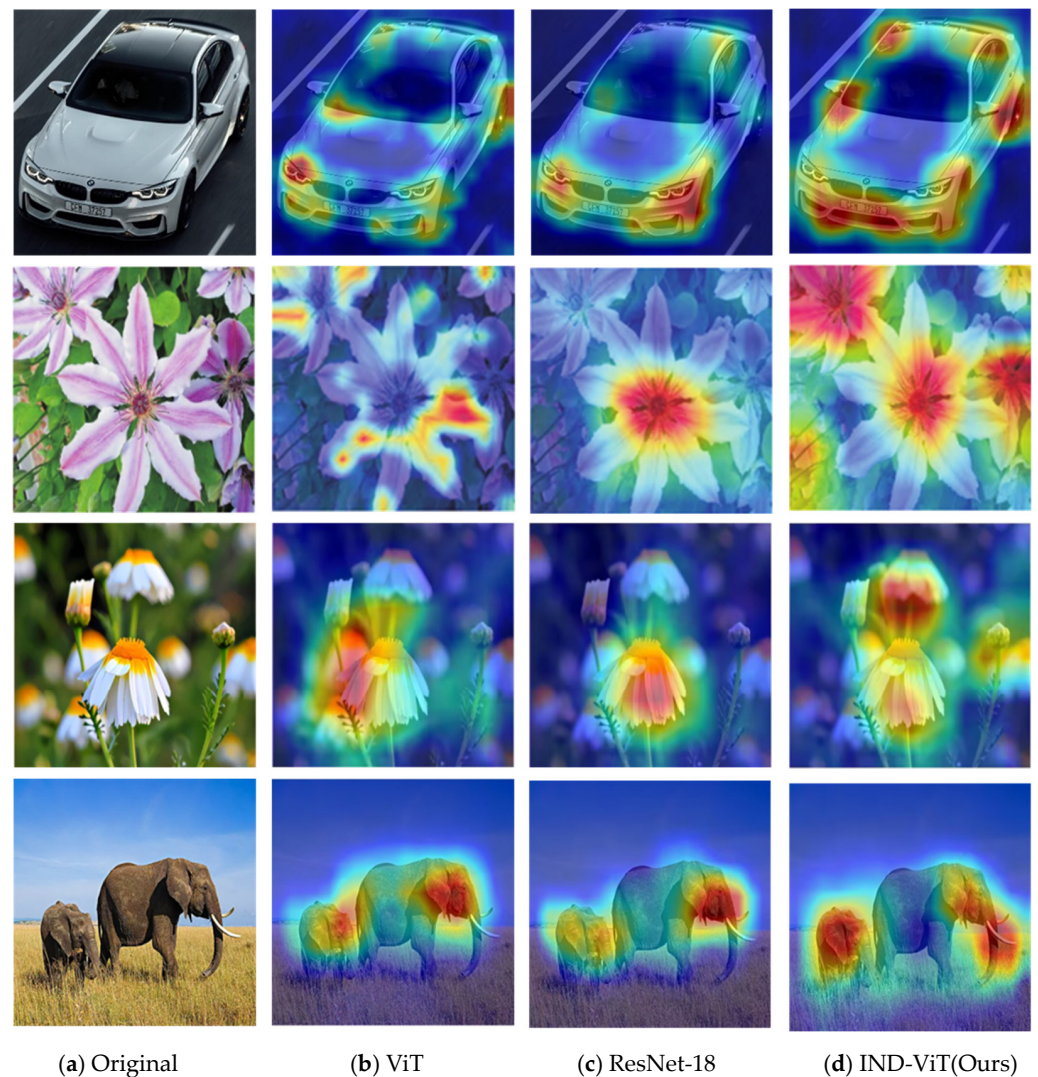


Figure 9. Visual comparison of heat maps.

4.6. Ablation Experiment

We carried out corresponding ablation experiments to analyze the effects of different modules on vehicle image recognition and evaluated the effects of the local detail feature extraction module, sparse attention module, and contrast loss, respectively.

In order to verify the validity and rationality of each algorithm module, ablation experiments were conducted on the local feature extraction module (CNN-In D) and sparse attention module (SAM). Ablation experiments were carried out in BIT-Vehicles, CIFAR-10, Oxford Floral-102, and Caltech-101 validation sets. The average accuracy of image classification was used as a measure of algorithm performance, and the algorithm was evaluated based on the pre-trained network architecture. The evaluation results are shown in Table 4, and the optimal value in each experiment is shown in bold. The column

Exp. represents the experiment number, and CNN-In D represents whether to use the local feature information extraction module proposed in Section 3.1. The column SAM represents whether to use the sparse attention module proposed in Section 3.4 to improve the image identification region recognition ability; Exp.1 is the original ViT model.

Table 4. Contribution of each module to the network.

Exp	CNN-In D	SAM	BIT-Vehicles	CIFAR-10	OxfordFlowers-102	Caltech-101
1	--	--	0.9428	0.9512	0.7358	0.6475
2	✓	--	0.9639	0.9684	0.7537	0.6518
3	--	✓	0.9785	0.9626	0.7451	0.6630
4	✓	✓	0.9842	0.9961	0.7742	0.6854

Local Detail Feature Extraction Module (CNN-In D): Table 4 shows the comparison of the average accuracy of different data sets when different modules are added. As can be seen from Table 4, the local feature extraction module proposed in Section 3.1 (CNN-In D) was added to the baseline model in the Exp.2 experiment, which significantly improved the performance of the model. The accuracy of the three data sets is improved by 2.11%, 1.72%, 1.79%, and 0.43%, respectively. This shows that the addition of the local detail feature extraction module makes up for the local detail features lost by the vision transformer in the direct partitioning operation of the image, which greatly improves the performance of the classification network.

Sparse Attention Module (SAM): As shown in Table 4, the sparse attention module (SAM) proposed in Section 3.4 of Exp.1 is added to Exp.3 to improve the recognition of similar images by the network. Through the sparse attention module, significant image blocks are selected as the input of the final coding layer. Compared with the original vision transformer model, the accuracy of our model on the four datasets is improved by 3.57%, 1.14%, 0.93%, and 1.55%, respectively, and the most significant is on the vehicle image classification dataset, which proves the effectiveness of the sparse attention module in vehicle classification. In addition, the Exp.4 experiment also adds a sparse attention module (SAM) to Exp.2. The recognition accuracy of the model can be improved from 96.39%, 96.84%, 75.37%, and 65.18% to 98.42%, 99.21%, 77.42%, and 68.54%, respectively. We believe that through the method of sparse attention, the model will sample the most discriminating image blocks as input so as to explicitly discard some of the most useless image blocks and force the network to learn the important parts, make full use of this weight information to realize the localization of the recognizable region, and improve the image classification ability.

Contrast Loss Function: Table 5 shows the recognition performance of the vision transformer and the model in this paper with and without a comparison loss function. The experiments were completed in the BIT-Vehicles, CIFAR-10, Oxford Flowers-102, and Caltech-101 validation sets, and the average accuracy of image classification was used as the metric of the algorithm performance, and the algorithm was evaluated based on the pre-trained network architecture. As shown in Table 5, the addition of contrast loss makes the vision transformer improve the recognition accuracy of the four data sets by 0.85%, 0.52%, 1.47%, and 0.31%, respectively, and the recognition accuracy of the model in this paper improves by 0.95%, 0.54%, 2.03%, and 1.90%, respectively. This proves that the comparison loss function is added in this paper, so that the similarity of the classification features corresponding to different labels is minimized and the similarity of the classification features of the same label is the largest, which improves the classification recognition accuracy and verifies the rationality of the loss function proposed by us.

Table 5. Ablation study of the contrast loss function.

Method	Contrast Loss	BIT-Vehicles	CIFAR-10	OxfordFlowers-102	Caltech-101
ViT	×	0.9535	0.9842	0.7267	0.6859
ViT	✓	0.9620	0.9894	0.7414	0.6890
IND-ViT	×	0.9645	0.9908	0.7834	0.7066
IND-ViT	✓	0.9740	0.9962	0.8037	0.7256

5. Conclusions

In order to solve the problem that the image segmentation operation in a vision transformer can easily lead to the loss of local details, this paper proposes a vehicle classification and recognition network structure based on an improved vision transformer. We designed CNN-In D, a local detail information extraction module, to make up for the local features lost by vision transformer networks and improve the perception capability of the networks. At the same time, a sparse attention module is proposed, which comprehensively uses the attention weight information of all coding layers to capture the discernible region in the image, and further improves the recognition ability of the model for similar vehicles. In addition, the contrast loss function is used to further increase the intra-class consistency and inter-class difference of network learning features. A large number of experiments show that both qualitative and quantitative visualization results demonstrate the effectiveness and interpretability of the proposed method. Compared with other recognition methods, the proposed method has higher recognition accuracy.

6. Discussion and Outlook

The vehicle classification algorithm (IND-ViT) based on the improved vision transformer proposed in this paper has achieved significant performance improvement in the vehicle classification task. Although the local detail feature extraction module (CNN-In D module) has shown results in extracting local features, there is still room for optimization. Future research can be devoted to the development of diverse feature extraction strategies to capture local details in images more comprehensively. In addition, although the sparse attention module enhances the model's ability to identify key regions, the further improvement of the attention mechanism is also worth exploring. In the future, we can study the method of adaptively adjusting the attention weight to improve the flexibility of the model to respond to the characteristics of different types of vehicles. The contrastive loss function introduced in this paper plays a key role in enhancing feature discrimination, and future research can explore the combination of other types of loss functions, such as triplet loss or center loss, to improve the generalization ability and robustness of the model. At present, this method is mainly aimed at vehicle classification tasks. Looking ahead, the research can be extended to the transferability of the model between different domains and tasks, such as adapting the vehicle classification model to other object classification tasks or enhancing the adaptability of the model in different environments. In the context of autonomous driving systems, deep learning models may be threatened by adversarial attacks. Therefore, future work can focus on improving the robustness of the model to such attacks and using techniques such as adversarial training and stochastic smoothing to enhance the resistance of the model. Finally, considering the multimodal data that may be involved in vehicle classification tasks, future research can explore cross-modal learning strategies to achieve more comprehensive vehicle recognition so as to promote the development of intelligent networked vehicle systems.

Author Contributions: Conceptualization, X.D. and P.S.; data curation, X.D. and L.Y.; formal analysis, X.D. and Y.T.; investigation, P.S.; methodology, X.D. and P.S.; project administration, A.Y.; software, L.Y.; visualization, X.D., Y.T. and T.L.; writing—original draft, X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (grant number: 2023CSJGG1600), Natural Science Foundation of Anhui Province (grant number: 2208085MF173), Wuhu “ChiZhu Light” Major Science and Technology Project (grant number: 2023ZD01) (grant number: 2023ZD03).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: Taonian Liang is an employee of Chery New Energy Automobile Co., Ltd. The paper reflects the views of the scientists, and not the company.

References

1. Won, M. Intelligent traffic monitoring systems for vehicle classification: A survey. *IEEE Access* **2020**, *8*, 73340–73358. [[CrossRef](#)]
2. Wang, P.; Ouyang, T.; Zhao, S.; Wang, X.; Ni, Z.; Fan, Y. Intelligent Vehicle Formation System Based on Information Interaction. *World Electr. Veh. J.* **2024**, *15*, 252. [[CrossRef](#)]
3. Dai, Z.; Guan, Z.; Chen, Q.; Xu, Y.; Sun, F. Enhanced Object Detection in Autonomous Vehicles through LiDAR—Camera Sensor Fusion. *World Electr. Veh. J.* **2024**, *15*, 297. [[CrossRef](#)]
4. Shi, D.; Chu, F.; Cai, Q.; Wang, Z.; Lv, Z.; Wang, J. Research on a Path Tracking Control Strategy for Autonomous Vehicles Based on State Parameter Identification. *World Electr. Veh. J.* **2024**, *15*, 295. [[CrossRef](#)]
5. Ressi, D.; Romanello, R.; Piazza, C.; Rossi, S. AI-enhanced blockchain technology: A review of advancements and opportunities. *J. Netw. Comput. Appl.* **2024**, *225*, 103858. [[CrossRef](#)]
6. Chen, Z.; Pears, N.; Freeman, M.; Austin, J. Road vehicle classification using support vector machines. In Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, Shanghai, China, 20–22 November 2009; Volume 4, pp. 214–218.
7. Kafai, M.; Bhanu, B. Dynamic Bayesian networks for vehicle classification in video. *IEEE Trans. Ind. Inform.* **2011**, *8*, 100–109. [[CrossRef](#)]
8. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [[CrossRef](#)]
9. Kaur, D.; Uslu, S.; Rittichier, K.J.; Durresi, A. Trustworthy artificial intelligence: A review. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–38. [[CrossRef](#)]
10. Butt, M.A.; Khattak, A.M.; Shafique, S.; Hayat, B.; Abid, S.; Kim, K.I.; Ayub, M.W.; Sajid, A.; Adnan, A. Convolutional neural network based vehicle classification in adverse illuminous conditions for intelligent transportation systems. *Complexity* **2021**, *2021*, 6644861. [[CrossRef](#)]
11. Deshpande, S.; Muron, W.; Cai, Y. Vehicle classification. In *Computer Vision and Imaging in Intelligent Transportation Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2017; pp. 47–79.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2017; p. 30.
13. 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 16×16 words: Transformers for image recognition at scale. *arXiv* **2010**, arXiv:11929.2020.
14. Maungmai, W.; Nuthong, C. Vehicle classification with deep learning. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019; pp. 294–298.
15. Yu, S.; Wu, Y.; Li, W.; Song, Z.; Zeng, W. A model for fine-grained vehicle classification based on deep learning. *Neurocomputing* **2017**, *257*, 97–103. [[CrossRef](#)]
16. Ma, Z.; Chang, D.; Xie, J.; Ding, Y.; Wen, S.; Li, X.; Si, Z.; Guo, J. Fine-grained vehicle classification with channel max pooling modified CNNs. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3224–3233. [[CrossRef](#)]
17. Jo, S.Y.; Ahn, N.; Lee, Y.; Kang, S.J. Transfer learning-based vehicle classification. In Proceedings of the 2018 International SoC Design Conference (ISODC), Daegu, Republic of Korea, 12–15 November 2018; pp. 127–128.
18. Neupane, B.; Horanont, T.; Aryal, J. Real-time vehicle classification and tracking using a transfer learning-improved deep learning network. *Sensors* **2022**, *22*, 3813. [[CrossRef](#)] [[PubMed](#)]
19. Hasanvand, M.; Nooshyar, M.; Moharamkhani, E.; Selyari, A. Machine learning methodology for identifying vehicles using image processing. *Artif. Intell. Appl.* **2023**, *1*, 170–178. [[CrossRef](#)]
20. Zhao, T.; He, J.; Lv, J.; Min, D.; Wei, Y. A comprehensive implementation of road surface classification for vehicle driving assistance: Dataset, models, and deployment. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 8361–8370. [[CrossRef](#)]
21. Zhao, D.; Chen, Y.; Lv, L. Deep reinforcement learning with visual attention for vehicle classification. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *9*, 356–367. [[CrossRef](#)]
22. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
23. Zhu, J.; Fang, L.; Ghamisi, P. Deformable convolutional neural networks for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1254–1258. [[CrossRef](#)]

24. Chen, Z.; Xie, L.; Niu, J.; Liu, X.; Wei, L.; Tian, Q. Visformer: The vision-friendly transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 589–598.
25. d’Ascoli, S.; Touvron, H.; Leavitt, M.L.; Morcos, A.S.; Biondi, G.; Sagun, L. Convit: Improving vision transformers with soft convolutional inductive biases. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 2286–2296.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
27. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 367–376.
28. Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. Levit: A vision transformer in convnet’s clothing for faster inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12259–12269.
29. Deshmukh, P.; Satyanarayana, G.S.R.; Majhi, S.; Sahoo, U.K.; Das, S.K. Swin transformer based vehicle detection in undisciplined traffic environment. *Expert Syst. Appl.* **2023**, *213*, 118992. [[CrossRef](#)]
30. Roy, S.K.; Deria, A.; Shah, C.; Haut, J.M.; Du, Q.; Plaza, A. Spectral-spatial morphological attention transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [[CrossRef](#)]
31. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
32. Loshchilov, I.; Hutter, F. Fixing weight decay regularization in adam. *arXiv* **2018**, arXiv:1711.05101.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
34. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
35. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

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