2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 978-1-6654-1865-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/MIPR51284.2021.00025

Transformer based Neural Network for Fine-Grained Classification of Vehicle Color

1st Yingjin Wang, 2nd Chuanming Wang, 3rd Yuchao Zheng, 4th Huiyuan Fu, 5th Huadong Ma
Beijing Key Lab of Intelligent Telecommunication Software and Multimedia
Beijing University of Posts and Telecommunications
Beijing 100876, China
{yingjin.wyj, wcm, yuchao.zyc, fhy, mhd}@bupt.edu.cn

***Abstract*—The development of vehicle color recognition technology is of great significance for vehicle identification and the development of the intelligent transportation system. However, the small variety of colors and the influence of the illumination in the environment make fine-grained vehicle color recognition a challenge task. Insufficient training data and small color categories in previous datasets causes the low recognition accuracy and the inflexibility of practical using. Meanwhile, the inefficient feature learning also leads to poor recognition performance of the previous methods. Therefore, we collect a rear shooting dataset from vehicle bayonet monitoring for fine-grained vehicle color recognition. Its images can be divided into 11 main-categories and 75 color subcategories according to the proposed labeling algorithm which can eliminate the influence of illumination and assign the color annotation for each image. We propose a novel recognition model which can effectively identify the vehicle colors. We skillfully interpolate the Transformer into recognition model to enhance the feature learning capacity of conventional neural networks, and specially design a hierarchical loss function through in-depth analysis of the proposed dataset. We evaluate the designed recognition model on the dataset and it can achieve accuracy of 97.77%, which is superior to the traditional approaches.**

***Index Terms*—Vehicle color recognition, Transformer, Self-attention, convolutional neural network**

I. Introduction

In recent years, the concept of Intelligent transportation system (ITS) has been proposed. ITS is the next generation of the transportation system which combines intelligent technology with existing transportation systems, and it has shown excellent performance in detecting violation of road rules and other routine cases. Vehicle, as one of the most important roles in ITS, has attracted the attention of more and more researchers.

Among the research attributes of vehicles such as color, model and license plate, vehicle color is regarded as a more valuable attribute than others. Because of its large proportion in vehicle body, strong anti-interference performance (blur, occlusion or visual angle change) and low recognition speciality, it is suitable for criminal cases or long-term vehicle video tracking tasks. If the intelligent monitoring platform can accurately and automatically identify vehicle color as fine-grained as possible, it can provide great help to crime investigation and vehicle tracking.

However, the present development of vehicle color recognition is mainly limited by the following aspects. First, the restricted categories in previous vehicle color datasets restrain the development of color recognition technology, in which the recognized vehicle colors are limited to 5 to 10 primitive colors. But in reality, fine-grained colors, such as purplish red and pink, which belong to the same primitive color, will have great differences under natural light, and it will bring great instability to the application of color recognition technology. Second, the color of vehicles in monitoring situations will be affected by interference factors such as different light conditions and weathers. To handle this issue, pretreatment techniques, such as artificial feature design, is usually used to carry out color screening and normalization of RGB values, whereas, in complex realistic environments, the above method can only achieve limited effect and can not be end-to-end used. Third, the previous methods don't take the global relation into consideration which is significant for fine-grained vehicle color recognition. The color of a vehicle is defined as the main color of its surface, and there will be some other colors that do not have the same RGB with its main color. Therefore, the recognition model needs to take in the global information and find the relationship between different regions of the vehicle.

To solve the above problems, we collect a rear shooting dataset for bayonet monitoring of vehicle color, which includes 32,220 vehicle images and covers different scenes and environmental changes. To adapt to the characteristics of the dataset, a priori labeling method using vehicle models is adopted, which can eliminate the influence of illumination. According to this method, we divide the images of the datasets into 11 color main categories and 75 color subcategories. To learn the features of the vehicle color globally, we propose a Transformer [1] based CNN model for vehicle color recognition. The Transformer modules are introduced to combine the traditional CNN, and make use of the self-attention mechanism to capture the key points of differences of the inputting feature maps. Besides, by referring to the idea of multi-scale feature fusion in FPN [2], the outputs of the last several layers of Transformer Encoder are concatenated to realize the combination of features in superficial and deep layers. To take the advantage of the dataset, we design a unique hierarchical loss function according to the two-level directory of vehicle

978-1-6654-1865-2/21/$31.00 ©2021 IEEE
DOI 10.1109/MIPR51284.2021.00025

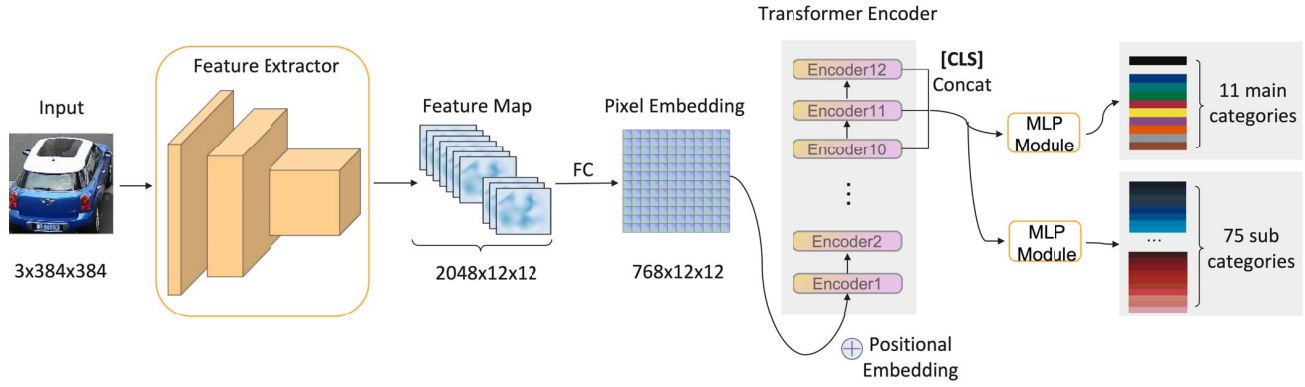

Fig. 1: The overall framework of our proposed approach. The modules are divided into Feature Extractor, Transformer Encoder and MLP multi-class classification modules.

color, which effectively improves the recognition precision. We evaluate our proposed method on the collected dataset via comparing many conventional CNN models and fine-grained recognition methods. The major contributions and innovation points of this paper are as follows:

- We collect a rear shooting dataset for vehicles and it includes 32,220 vehicle images and can be divided into 11 color main categories and 75 color subcategories by our proposed labeling method.
- We propose a Transformer based CNN model for fine-grained vehicle color recognition, which can make the model learn global feature maps and design a special loss for better supervision.
- We conduct experiments on the collected dataset, and the comparison with several CNN models and fine-grained recognition methods can demonstrate the effectiveness of our proposed approach.

The rest part of our paper is arranged as follows: we briefly review the research progress of vehicle color recognition, fine-grained classification and Transformer in Section II. The Section III and the Section IV introduces the collected dataset and the main architecture of our proposed method respectively. The description of the details of the experiment are presented in Section V. Finally, we do the acknowledgement and summarize our work in Section VI and Section VII.

II. Related Work

A. vehicle Color Classification

In terms of vehicle color classification, the existing work can be roughly divided into two types: the methods based on artificial features and the CNN method based on automatic features. Before the large-scale introduction of neural networks, the method of artificial feature extraction and then combination with classifiers occupied the dominant position, such as the common color histogram. Li et al. [3] use HSI color space as well as the matching algorithm of relative error distance for vehicle color recognition; Yang et al. [4] propose a new vehicle color recognition method by using HS histogram. Compared with the previous method, the CNN method has stronger generalization ability and higher accuracy. Rachmadi et al. [5] prove that CNN could be used for vehicle color recognition; Hu et al. [6] propose a vehicle color recognition algorithm based on deep learning. They combine the spatial pyramid strategy with the original CNN, which can adaptively learn the color features of vehicle; in recent years, the development of deep neural network is combined with the task of vehicle color classification (e.g., Kim et al. [7], Wu et al. [8], Fu et al. [9]) to further enhance the recognition precision of vehicle color classification.

B. Fine-grained Classification

For the research of fine-grained classification, the general deep convolutional neural network(D-CNN) method can hardly capture discriminatory local details. Other common and effective methods can be divided into the following three types: first is the method based on local labeling information, which relies on additional component-level intensive labeling information to locate key points, such as head, mouth, etc. (Lin et al. [10], Wei et al. [11]). Another method is to use attention to locate highly discriminatory regions. This method only needs image-level labeling information to realize accurate local positioning (Fu et al. [12], Zheng et al. [13], Sun et al. [14]). The above two methods first find out the discriminatory parts, and then extracts and classifies the features. Recently, some work shows that (Kong et al. [15], Cui et al. [16], Dubey et al. [17]) feature extraction can be directly completed end-to-end by developing powerful deep network, thereby learning more discriminatory feature representation.

C. Transformer in Computer Vision

Transformer was first proposed by Vaswani et al. [1] in 2017 for machine translation tasks, and uses the multi-headed self-attention mechanism to calculate the weight of each position in the global scope. Because of the outstanding global capture capability of that, the application of Transformer in computer

TABLE I: Color subcategories Contained in each main category

white	Cream	Greey white	Papyrus white	Pure white	Signal white	Traffic white	
grey	Light grey	Platinum grey	Signal grey	Silver grey	Squirrel grey	Tarpaulin grey	Yellow grey
black	Jet black	Traffic black					
brown	Chocolate brown Orange brown	Grey brown Signal brown	Pale brown	Red brown	Terra brown	Copper brown	Ochre brown
blue	Black blue Sky blue	Brilliant blue Steel blue	Distant blue Ultramarine blue	Gray blue	Light blue	Night blue	Sapphire blue
yellow	Beige	Ochre yellow	Saffron yellow	Sand yellow	Sulfur yellow	Ivory	Lemon yellow
red	Antique pink Purple red	Beige red Raspberry red	Black red Ruby red	Coral red Signal red	Flame red Strawberry red	Luminouse red Light pink	Orient red Traffic red
green	Black green	Olive green	Pastel green	Reed green	Reseda green	Yellow green	
purple	Blue lilac	Pearl violet	Red violet	Heather violet	Telemagenta		
orange	Pastel orange	Pearl orange	Salmon orange	Signal orange	Luminous orange		
cyan	Light green	Pastel turquoise	Azure blue	Mint turquoise			

vision is a hot research trend recently. However, limited by the square-level computational expense of pixel number, it is unrealistic to directly apply Transformer to the original image. Given this, there are two main solutions: first is to use an approximation method. For example, Dosovitskiy proposes the VIT model [18] in the latest research, which divides the input image into 16×16 patches, and directly feed these patches into the standard Transformer. The model structure is simple, but due to the lack of local correlation, mass of data are needed to carry out long-time pre-training; second is to combine CNN with Transformer. Usually, Transformer is used to further process the output of CNN network, such as in image recognition (Dert [19]) and image classification (Wu et al. [20]). Such a structure can use the Transformer to achieve global capture capability, as well as reserve the CNN structure to prevent the translation invariance from being destroyed. However, to avoid the high computational cost brought by Transformer, complex and redundant improvements have been made, which cannot directly reflect the value of Transformer for image information extraction.

III. Dataset

To cover the common vehicle colors as far as possible, a brand new dataset of 32,220 vehicle images, including 11 main categories and 75 subcategories, is collected, as shown in Table I and Fig.2. The vehicle images are shot from rear view by road HD bayonet camera, and each image contains only one complete vehicle. In addition, the dataset covers multiple models such as family cars, sports cars, buses and trucks, which ensures the diversity and completeness of the dataset.

The prior labeling algorithm. The collected images required to label the class manually, therefore, in this section, a prior labeling algorithm for our dataset that can eliminate the influence of illumination is proposed. In the labeling process of our data, because of the large color deviation under different light and weather conditions, and the small class difference in fine-grained classification, it is difficult to judge the real vehicle color and give an accurate label by human eyes. In view of this, we adopt the prior information of vehicle models, that is, for the same series of vehicles of the same brand, the color classes are definite and limited, and all vehicles in each coarse-grained color class belong to the same fine-grained color. The detailed steps are described as follows.

First, we divide all images into groups by the fine-grained models and coarse-grained vehicle colors. Second, for the images with no obvious abnormal light and weather, which belong to the same fine-grained model and coarse-grained vehicle color, we extract RGB values of all pixels for each image and put these values into a set S. Third, for set S, K-means [21] method is used to carry out clustering (K = 5), the central point of each cluster corresponds to a main color of the specific vehicle type, then we get the top-5 RGB values as the fine-grained colors of the specific vehicle model with the same coarse-grained color. Fourth, we visualize the five main fine-grained colors obtained by clustering, and the best appropriate RGB value of the top-5 RGB values of the specific model and color is manually screened out by comparing the clustered colors and the appearance of the vehicle.

Finally, we need to map the selected RGB values into one of our 75 color categories. According to Equation 1, which reflects the calculation process of RGB distances, the category with the smallest distance, that is, the label of all the vehicle images of the specific model and color, are selected. By using the above method, the present dataset can cover images that are shot under extreme weather and extreme light with accurate label information, which ensures the completeness of the dataset.

$$\bar{C} = \sqrt{(2 + \frac{\bar{r}}{256}) \times \bar{R}^2 + 4 \times \bar{G}^2 + \frac{767 - \bar{r}}{256} \times \bar{B}^2} \quad (1)$$

where $\bar{C}$ means the distance between the RGB value obtained in the fourth step and each RGB value in the fine-grained color class standard, and $\bar{R}$, $\bar{G}$, $\bar{B}$ are obtained by $\bar{X} = C_{1,X} - C_{2,X}$, where X can be {R, G, B}, and $\bar{r}$ is a

value related to R, the calculation process can be formulated as $\bar{r} = \frac{C_{1,R}+C_{2,R}}{2}$. C_1 and C_2 are the two colors being calculated.

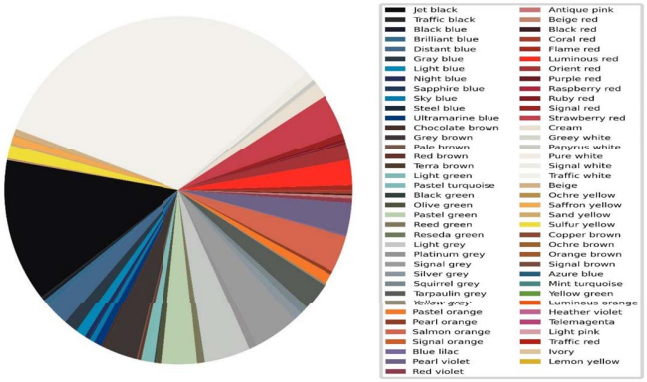

Fig. 2: The proportion of images of each color subcategory in the dataset.

IV. Methodology

The architecture of our proposed model is shown in Fig.1. First, the input images are processed by the feature extractor which can be pre-trained on the Large-scale dataset for transfer learning, and then the output feature maps are fed to the Transformer. Meanwhile, inspired by the idea in FPN [2], we adopt a multi-scale feature fusion to concatenate the outputs of the last three layers of Transformer Encoder, thereby further realizing the feature fusion of deep and superficial layers. In the optimization stage, we design an adaptive hierarchical loss function based on the two-level directory of vehicle color.

About the feature extractor in our model, we use the ResNet26-D [22] [23] pre-trained on ImageNet [24] to extract the shallow features F of the input images. Then, to meet the requirements that the input data of Transformer Encoder received must be an one-dimensional token embedding sequence, we segment the generated feature map F in pixels as pixel embedding, and use the fully connected network to adjust channels of per pixel embedding to D for further fusing features between channels and reduce the amount of parameters in following Transformer Encoder. Finally, the number of all the obtained pixels M is the effective input sequence length of Transformer Encoder.

Now, we have obtained the pixel embedding sequence, which the dimension size is $M \times D$. Besides, to reserve the position information of the feature map, the input embedding of Transformer also needs to cover position embedding, and we adopt the randomly initialized and learnable 1D position embedding E_{pos} in our model. And to perform the subsequent classification task, we need a stable token [CLS] to represent the features learned by all pixels, then feed that token to the classification network. Therefore, the input of the first encoder in Transformer can be formulated as Equation 2:

$$Z_0 = [X_{[CLS]}; X_{feature\_map}] + E_{pos}, \tag{2}$$

where $X_{[CLS]} \in \mathbb{R}^{1\times D}$, $X_{feature\_map} \in \mathbb{R}^{N\times D}$, and $E_{pos} \in \mathbb{R}^{(N+1)\times D}$. [CLS] itself has no semantics, through 12 layers of Encoder, it will integrate the feature information in each pixel more fairly. Transformer works in an iterative manner, and the output of the last encoder will be sent to the next encoder as its input. We formulate this process as Equation 3:

$$\begin{aligned} Z_i^* &= \mathrm{MultiHeadAtt}(Z_{i-1}, Z_{i-1}, Z_{i-1}) + Z_{i-1} \\ Z_i &= \mathrm{FFN}(\mathrm{LN}(Z_i^*)) + Z_i^*, \end{aligned} \tag{3}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{D\times(D/h)}$.

Each layer is composed of the MultiHeadAtt part connecting to the FFN part (double-layer linear transformation and GELU [25] activation function). Residual connection and Layer Normalization(LN) [26] is used after every block. The specific implementation details of this part are shown in Fig.3, and the MultiHeadAtt is formulated as Equation 4:

$$\begin{aligned} &\mathrm{MultiHeadAtt}(Q,K,V) = \mathrm{Concat}(head_1, ..., head_h) \\ &where\ head_i = \mathrm{Softmax}(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}})(VW_i^V) \end{aligned} \tag{4}$$

For FFN, by using two-layer linear transformation and GELU activation function, it can do the linear integration and combine the features on different channels. It can be formulated as Equation 5:

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2, \tag{5}$$

where $W_1 \in \mathbb{R}^{D\times 4D}$, $W_2 \in \mathbb{R}^{4D\times D}$.

When applying the traditional Transformer encoder, we only feed the output [CLS] token of the last encoder layer into the MLP module. To further fuse the deep features with superficial features, the idea of multi-scale feature fusion inspired by FPN is used to concatenate the output of the [CLS] token of the last three layers in Transformer Encoder, which is then input into the subsequent classification network after the adjustment of dimension. Compared with the original Transformer encoder, which only uses the output of the last layer, the recognition accuracy is further improved.

The recognition accuracy of main color categories is higher due to its obvious difference than the fine-grained colors. Therefore, considering the recognition of main categories in back propagation can assist the recognition of fine-grained colors, we design an adaptive hierarchical loss which follows the idea of hierarchical classification. It enforces the recognition model first distinguish the main color categories, and then subdivide the subcategories covered in each main category. We combine the losses of main and fine-grained classes as the joint loss to supervise the learning of the recognition model. The specific calculation process is shown as Equation 6:

$$\mathrm{Loss} = \mathrm{CE}(pred1, label1) + \mathrm{CE}(pred2, label2), \tag{6}$$

Where CE refers to CrossEntropyLoss. $label1$ is the main category corresponding to the current image, which is an integer between 0 and 10, and $pred1$ is the value of each main category the present image belongs to, which is predicted by the model, and the size is 1×11. After the Softmax function,

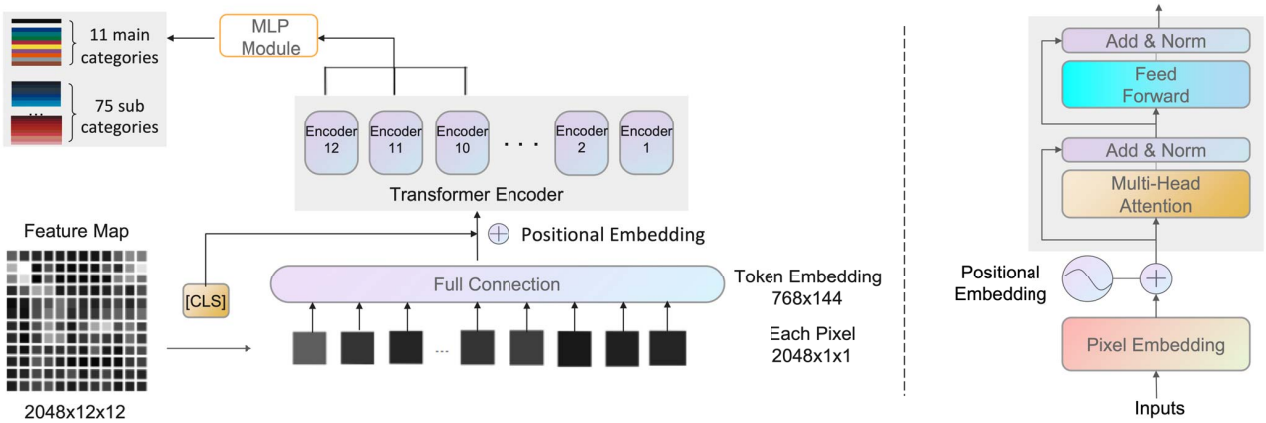

Fig. 3: The implementation details of Transformer are shown on the left. The generated feature map is divided into pixel blocks with fixed size. To perform the subsequent classification task, a stable and learnable [CLS] token is added, which has no semantics and integrates the feature information in each pixel. And position embeddings are added to reserve the position information of the feature map. Then the pixel embedding sequence is input into the Transformer Encoder. On the right is the illustration of Transformer Encoder inspired by Vaswani et al. [1], and we stack 12 layers of that in our model.

the probability of the present image belonging to each main category can be obtained, which is between 0 and 1. $label2$ is the subcategory of the present image, which is an integer between 0 and 75. Similarly, $pred2$ is the corresponding value of each subcategory, and its size is 1×75.

V. Experiment

We propose a new dataset containing dozens of vehicle colors. Based on it, experiments are conducted to evaluate the performance of the proposed algorithm and compare it with other CNN models and fine-grained recognition methods.

A. Implementation Details

We adopt the implementation of PyTorch [27] to train the present model, and one NVIDIA 2080 GPU is used for training. In terms of parameter setting, we set the learning rate to 0.0003 and maintain it for all the training process. We set the epochs to 100, SGD as the optimizer, the momentum to 0.9, and the weight decay to 0.0001.

The dataset is randomly divided into the training set and validation set according to the ratio of 4:1. The input image has been adjusted to $3\times384\times384$ for feature extraction. After ResNet26-D, the output feature map has the shape of $2048\times12\times12$. Then, to meet the input requirements of the Transformer Encoder, we use a fully-connected network to transfer the feature map before entering the Transformer Encoder module. This network maps each element in the feature map ($2048\times1\times1$) into a vector of fixed dimension D (D=768). As described in Section IV we add each vector together with a corresponding randomly initialized position embedding. The generated 145 (including [CLS] token)$\times$768 two-dimensional vector diagram is then input into the Transformer Encoder, and the self-attention mechanism is used to focus on the key parts in the vehicle image. The [CLS] token vectors in the last three layers of the Transformer are obtained and concatenated together to output a two-dimensional vector as the feature of the input image, which has a shape of 3×768. Finally, a fully-connected network is followed by the Transformer Encoder for multi-class classification, and it will classify the input image into 11 color main categories and 75 color subcategories depending on its extracted feature.

B. Experimental Results

In our model, We use the combination of basic pre-trained CNN with Transformer. As shown in Table II and Table III, it can be found that compared with traditional classical model and common fine-grained classification model, the recognition accuracy of the present method has been greatly improved. And the recognition accuracy of each color subcategory with more than 10 pictures is shown in Fig.4.

About the model compared in our paper, such as traditional classical model-ResNet [22], VGG [28] and SeNet [29]. ResNet is based on residual connection to endow the backpropagation gradient with better local correlation, VGG is based on AlexNet [30] network to further study the depth and width of deep neural network, SeNet is based on the idea of attention to learn the importance of each channel. It is found through comparison that among various models, our method performs best in the present task. See Table II for the details.

And about the two common fine-grained models being compared(BCNN [31] and DFL-CNN [32]). BCNN uses matrix outer product to combine the feature maps of two CNNs, and DFL-CNN learns the mid-level representation of the enhanced model through learning a bank of convolutional filters. Compared to the common fine-grained models, our

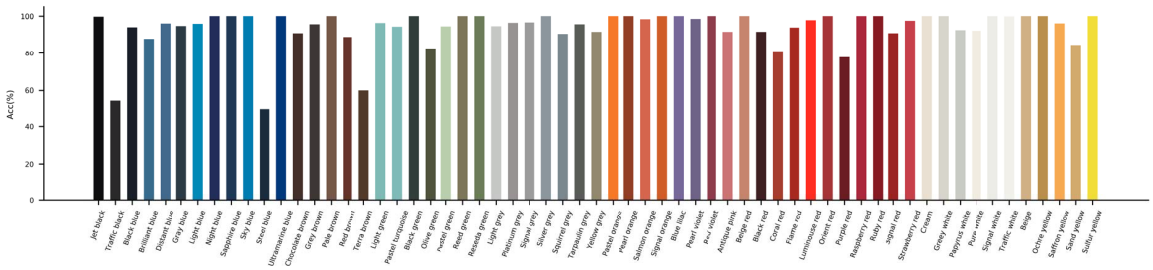

Fig. 4: Recognition accuracy of each color subcategory with more than 10 pictures

TABLE II: Comparison with various basic CNN Models

Method	Accuracy	△
ResNet-18	93.95%	-3.82%
ResNet-34	94.63%	-3.14%
SeNet-154	91.95%	-5.82%
VGG-16	93.36%	-4.41%
Ours	**97.77%**	-

method not only improves the accuracy, but also dramatically reduces computational cost. See Table III for the details.

TABLE III: Comparison with common Fine-Grained models

Method	Accuracy	△	FLOPs(G)
BCNN	88.18%	-9.59%	61.54
DFL-CNN	96.60%	-1.17%	62.84
Ours	**97.77%**	-	**20.2**

In addition, as shown in Table IV, to verify the effectiveness of each module, the ablation experiment is conducted. VIT model is used to cut the image into 16×16 image blocks, and then the features of each image block are extracted as embedding. However, this method does not pay sufficient attention to the local information of each position in the original image. Therefore, in the present method, first Resnet-26-D is used to fuse the image information. Meanwhile, based on the feature map, the outstanding global capture ability of Transformer can better capture the relationship between the pixels in the feature map. Besides, because different layers of Transformer Encoder capture different levels of semantic and syntactic information, the head structure is customized, the outputs of the last three layers of the encoder are selected for concatenating, and then connected to the MLP module, which improve the accuracy by 1.1%. Furthermore, the effectiveness of the user-defined hierarchical loss function is verified. Because of the great difference between the main color categories, and the recognition accuracy is higher, hence, the hierarchical loss function takes into account the predicted results of main color categories, which improves the accuracy by 0.3%.

TABLE IV: Ablation experiment result, which reflects the influence of different components.

Method	Accuracy
Resnet-26-D	94.03%
Transformer	87.91%
Resnet-26-D+Transformer	96.41%
Resnet-26-D+Transformer+FPN	97.50%
Resnet-26-D+Transformer+FPN+Hierarchical Loss(Ours)	**97.77%**

VI. Acknowledgement

This work was supported in part by the National Key R&D Program of China under No.2017YFB1003000, the Beijing Natural Science Foundation(L191004), the National Natural Science Foundation of China under No.61872047 and No.61720106007, the Beijing Nova Program under No.Z201100006820124, and the 111 Project (B18008).

VII. Conclusion

In this paper, we focus on the problem of fine-grained vehicle color recognition, and find that the previous technology is limited by the restricted dataset and ineffective learning methods. Therefore, we collect a vehicle dataset, which contains a total of 32,220 images, and divide the images into 11 main-categories and 75 subcategories. For this dataset, We propose a labeling algorithm which can eliminate the influence of illumination. And we propose a Transformer based Neural Network for fine-grained vehicle color recognition, which utilizes the global learning ability of Transformer, and we skillfully interpolate it with CNN model. Besides, we design an adaptive hierarchical loss function which considers the hierarchical structure of the dataset. The evaluation on the dataset demonstrates that our method can surpass other CNN models and fine-grained recognition methods, showing the effectiveness of our approach.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. I, II-C, 3

[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. I, IV

[3] X. Li, G. Zhang, J. Fang, J. Wu, and Z. Cui, "Vehicle color recognition using vector matching of template," in *2010 Third International Symposium on Electronic Commerce and Security*. IEEE, 2010, pp. 189–193. II-A

[4] M. Yang, G. Han, X. Li, X. Zhu, and L. Li, "Vehicle color recognition using monocular camera," in *2011 International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2011, pp. 1–5. II-A

[5] R. F. Rachmadi and I. Purnama, "Vehicle color recognition using convolutional neural network," *arXiv preprint arXiv:1510.07391*, 2015. II-A

[6] C. Hu, X. Bai, L. Qi, P. Chen, G. Xue, and L. Mei, "Vehicle color recognition with spatial pyramid deep learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2925–2934, 2015. II-A

[7] K.-J. Kim, P.-K. Kim, K.-T. Lim, Y.-S. Chung, Y.-J. Song, S. I. Lee, and D.-H. Choi, "Vehicle color recognition via representative color region extraction and convolutional neural network," in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2018, pp. 89–94. II-A

[8] X. Wu, S. Sun, N. Chen, M. Fu, and X. Hou, "Real-time vehicle color recognition based on yolo9000," in *International Conference in Communications, Signal Processing, and Systems*. Springer, 2018, pp. 82–89. II-A

[9] H. Fu, H. Ma, G. Wang, X. Zhang, and Y. Zhang, "Mcff-cnn: Multiscale comprehensive feature fusion convolutional neural network for vehicle color recognition based on residual learning," *Neurocomputing*, vol. 395, pp. 178–187, 2020. II-A

[10] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep lac: Deep localization, alignment and classification for fine-grained recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1666–1674. II-B

[11] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognition*, vol. 76, pp. 704–714, 2018. II-B

[12] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446. II-B

[13] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5209–5217. II-B

[14] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 805–821. II-B

[15] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 365–374. II-B

[16] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2921–2930. II-B

[17] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum-entropy fine grained classification," in *Advances in neural information processing systems*, 2018, pp. 637–647. II-B

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. II-C

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020. II-C

[20] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, M. Tomizuka, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020. II-C

[21] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297. III

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. IV, V-B

[23] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558–567. IV

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. IV

[25] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016. IV

[26] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016. IV

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037. V-A

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. V-B

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. V-B

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. V-B

[31] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457. V-B

[32] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a cnn for fine-grained recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4148–4157. V-B