Front. Comput. Sci., 2023, 17(3): 173321
https://doi.org/10.1007/s11704-022-1389-x

RESEARCH ARTICLE

Vehicle color recognition based on smooth modulation neural network with multi-scale feature fusion

Mingdi HU (✉)[1], Long BAI[1], Jiulun FAN[1], Sirui ZHAO[2], Enhong CHEN[2]

1 School of Communications and Information Engineering & School of Artificial Intelligence, Xi'an University of Posts & Telecommunications, Xi'an 710121, China

2 School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China

Abstract Vehicle Color Recognition (VCR) plays a vital role in intelligent traffic management and criminal investigation assistance. However, the existing vehicle color datasets only cover 13 classes, which can not meet the current actual demand. Besides, although lots of efforts are devoted to VCR, they suffer from the problem of class imbalance in datasets. To address these challenges, in this paper, we propose a novel VCR method based on Smooth Modulation Neural Network with Multi-Scale Feature Fusion (SMNN-MSFF). Specifically, to construct the benchmark of model training and evaluation, we first present a new VCR dataset with 24 vehicle classes, Vehicle Color-24, consisting of 10091 vehicle images from a 100-hour urban road surveillance video. Then, to tackle the problem of long-tail distribution and improve the recognition performance, we propose the SMNN-MSFF model with multi-scale feature fusion and smooth modulation. The former aims to extract feature information from local to global, and the latter could increase the loss of the images of tail class instances for training with class-imbalance. Finally, comprehensive experimental evaluation on Vehicle Color-24 and previously three representative datasets demonstrate that our proposed SMNN-MSFF outperformed state-of-the-art VCR methods. And extensive ablation studies also demonstrate that each module of our method is effective, especially, the smooth modulation efficiently help feature learning of the minority or tail classes. Vehicle Color-24 and the code of SMNN-MSFF are publicly available and can contact the author to obtain.

Keywords vehicle color recognition, benchmark dataset, multi-scale feature fusion, long-tail distribution, improved smooth l1 loss

1 Introduction

Statistically speaking, the number of motor vehicles globally has reached billions by 2021, which brings great challenges to intelligent traffic management and tracking vehicle crimes. At present, real-time vehicle recognition mainly depends on the license plates [1], but they often cannot work well if the camera fails to capture the correct license plates because of being changed, covered, removed, or forged. As another essential attribute [2], vehicle color can assist traffic law enforcement and improve the reliability of vehicle recognition in tracking crime vehicles, thus, recognizing vehicles automatically through vehicle color is paid more and more attention, and has gradually been applied in practice [3].

Recent years, there are lots of efforts devoted to vehicle color recognition, which mainly fall into two kinds: hand-crafted methods [3–6] and deep learning methods [7–11]. The hand-crafted methods, also termed traditional methods [3–6], usually first extract superficial visual features by using manual designed feature descriptors, then utilize supervised classifiers such as Support Vector Machines (SVM) to classify vehicles into 13 color categories at most. However, the hand-crafted feature descriptors rely on prior professional knowledge and are time-consuming, and the corresponding performance on a new task is also difficult to guarantee. For the deep-learning method, they usually pay more attention to constructing complex network models to improve recognition accuracy. Nowadays, some popular deep learning models for target recognition, e.g., Faster-RCNN [12], SSD [13], YOLO-v3 [14], YOLO-v4 [15], Efficient-Det [16], Center-Net [17], Retina-Net [18], also have been applied to vehicle color recognition [2,9,10] due to their superior performance, however, these methods rarely involve the imbalance distribution nature of datasets.

Actually, deep learning models are data-driven, and their performance is mainly determined by two key factors: neural network configuration and data corpus, especially the latter. The diversity of data types, sample size, accuracy of data annotation, rationality of data cleaning, and mining of prior knowledge of data distribution all contribute to improving the performance of the algorithm. However, the existing three public datasets all have drawbacks. The vehicle color dataset constructed by Chen et al. [3], for short C-dataset, which includes only 8 color categories, and each image contains only one vehicle. The dataset collected by Jeong et al. [4] is abbreviated as J-dataset, which includes only 7 colors, and it

Received July 16, 2021; accepted March 30, 2022

E-mail: humingdi2005@xupt.edu.cn

has a very small sample size. Tilakaratna et al. [5] attempted to expand color categories of the vehicle color dataset, but their dataset just reaches 13 categories, which is called T-dataset briefly.

To address the above challenges, we first establish a new benchmark dataset of vehicles with 24 colors, and then propose a novel VCR method based on Smooth Modulation Neural Network with Multi-Scale Feature Fusion (SMNN-MSFF). For the benchmark dataset, i.e., Vehicle Color-24, which includes 10091 images and each image contains 1–9 vehicles that amount to 31232. Besides, to improve the quality of our dataset and intuitive presentation, we conduct pre-processing operations and data analysis. The result of data analysis shows a natural tendency or long-tailed effect. For our SMNN-MSFF, considering the color information is a low-level visual feature, we first adopt a light-weight network with 42 CNN layers as our backbone. Then, considering the vehicle identification requires rich location information such as edges and corners, we add Feature Pyramid Networks (FPN) module in our framework that is inspired by the Faster-RCNN network structure, and multi-scale information fusion is implemented at the same time. Particularly, we use improved Smooth L1 loss function to eliminate the influence of the long-tailed effect [19]. The contributions of our method are threefold:

1) We build a new dataset with 24 vehicle colors, called Vehicle Color-24. The data was collected from the traffic management surveillance videos of the public security bureau. The colors of the Vehicle Color-24 are divided into 24 types, including red, dark-red, pink, orange, dark-orange, red-orange, yellow, lemon-yellow, earthy-yellow, green, dark-green, grass-green, cyan, blue, dark-blue, purple, black, white, silver-gray, gray, dark-gray, champagne, brown and dark-brown. To the best of our knowledge, Vehicle Color-24 is the largest VCR datasets, and can make up for the current needs of practical vehicle traffic management and criminal vehicle tracking applications.

2) We propose a novel vehicle color recognition method based on SMNN-MSFF. Firstly, our algorithm starts to pay attention the color distribution imbalance nature existing in any dataset. The loss function fine-tunes the network so that the algorithm can better capture the characteristics of small-scale classes than focal loss through ablation experiments. Secondly, our network adds FPN module to extract edges and corners information, which is helpful to extract vehicle shape features and local location information to assist vehicle recognition. Thirdly, our backbone network is designed with only 42 layers, which belongs to lightweight network, to relieve the pressure of storage and increase possibility of implementation in practical applications.

3) Comprehensive experiments demonstrated that performance of SMNN-MSFF has achieved significant improvement in classifying vehicle color.

The remainder of the paper is organized as follows: Section 2 introduces the related work. Section 3 describes the construction, analysis and pre-processing of the Vehicle Color-24 in detail. The overview and structure of every module of SMNN-MSFF are described in Section 4. The experimental design and result are shown in Section 5. And conclusions are presented in Section 6.

2 Related work

2.1 Vehicle color datasets

In general, a vehicle has only one kind of dominant color, and the vehicle color is a significant attribution in all characteristics of the entire vehicle and is not easy to be damaged or changed. Therefore, the vehicle color attribute based methods for target detection and recognition are worth to be exploited effectively. By 2014, Chen et al. [3] proposed a method to recognize vehicle color on urban roads by extracting feature context, and was the first to construct a public benchmark dataset for vehicle color recognition. The benchmark C-dataset contains two sub-datasets of images and videos, but includes only 8 colors with 15601 images, and each image contains only one vehicle. All categories such as white, black, gray, red, cyan, blue, yellow and green. Tilakaratna et al. [5] presented image analysis algorithms for vehicle color recognition, and the meanwhile established the T-dataset with 13 colors types, just as silver, white, black, gray, sky blue, red, blue, brown, orange, pink, green, yellow and purple. Images of T-dataset were collected using highway traffic cameras which are mounted on Thailand motor ways in day time. Jeong et al. [4] defined a new color homogeneity patch and constructed the J-dataset. They focused on improving the discernment of vehicle color features by avoiding another color distribution, e.g., saturated by day light, color of headlight, color of radiator grille, color gradation. The above three datasets are relatively large scale and have been widely used in main vehicle color recognition literature. Each of them has its own characteristics and focus. However, they contain only 13 color types at most, which is far insufficient to meet the requirements of fine color recognition of vehicles. Till now, few researchers pay attention to more vehicle colors, which poses a severe difficulty for the development of relevant research, so we construct a public benchmark dataset of vehicle colors with 24 types.

2.2 Algorithms of vehicle color recognition

The extensive application value of vehicle color recognition has attracted researchers in the field of computer vision, who have proposed many efficient recognition methods [3–11]. All existing main vehicle color recognition methods fall into two categories: manual feature-based traditional methods and deep learning-based data-driven methods. Chen et al. [3] took the representation of Feature Context (FC) [20] to describe the image and divide the image into fan-like subregions. Then they used based Bag of Word (BoW) method to extract color histogram features and map the color features into a higher dimensional subspace by quantizing them using a large codebook. And last they trained a linear SVM as their classifier. Jeong et al. [4] detected an Regions of Interest (ROI) of a vehicle image and the color of the vehicle was determined by integrated HSV histogram feature of each patch with the voting strategy. Dule et al. [6] examined two ROI (smooth hood peace and semi front vehicle), three classification methods (K Nearest Neighbors, Artificial Neural

Networks and SVM), and all possible combinations of sixteen color space components as different feature sets. Traditional methods basically adopt hand-crafted features design, which requires professional knowledge. An appropriate feature usually takes massive experience and time to validate its effectiveness for a certain task. In recent years, deep-learning based methods have successfully surpassed traditional manual feature methods for solving many visual tasks. There existed representative works in vehicle color recognition based on deep learning, for example, Hu et al. [7] proposed convolutional neural networks with spatial pyramids strategy, and applied SVM to divide vehicle color into 8 types. Rachmadi et al. [8] improved deep neural network introduce by Krizhevsky et al. [21], and converted the input image of C-dataset [3] to two different color spaces, HSV and CIE Lab, and then divided the vehicle color into 8 classes. They just classified 8 color types and the running time was long. Zhuo et al. [9] proposed a high-precision vehicle color recognition method for urban surveillance video based on a hierarchical fine-tuning strategy, but only 10 color categories are involved. Fu et al. [10] used a multi-scale feature fusion module based on the residual network for vehicle color recognition, and established a system for vehicle color recognition in actual traffic circumstances, which is pity that only 8 color types is considered like before.

2.3 Category imbalance

As the class size grows, maintaining a balanced dataset across many classes is challenging. All most the data are long-tailed in nature, so is vehicle color dataset. At present, the solutions to the category imbalance mainly include: Re-sampling and Re-weighting [18,22–26], Representation learning [27,28], Transfer learning [29–32], Logit adjustment [19,33], and Ensemble learning [34–36]. Common loss functions for target detection include: L1 loss, cross entropy loss (CE Loss), mean square error loss (MSE Loss) and focal loss [18]. And the Smooth L1 loss function can prevent gradient explosion and has the advantages of both CE loss and L1 loss. It can more flexibly adapt to the number of samples of different categories and cause different loss values by adding a parameter β. So we leveraged the Smooth L1 Loss to alleviate the class imbalance in the fine color recognition of vehicles, and further fine-tuning with other loss functions will be used to improve the identification accuracy of minority class.

2.4 Object detection

In particular, Faster-RCNN [12], SSD [13], YOLO-v3 [14], YOLO-v4 [15], Efficient-Det [16], Center-Net [17], Retina-Net [18], etc. have been widely applied to vehicle color recognition due to their outstanding detection performance [2,9,10]. However, these methods almost tend to design deeper and deeper structures with lots of parameters, which is prone to be over-fitting and long-running time. Therefore, in this paper, we designed a relatively light network SMNN-MSFF method with four parts with 64 layers.

3 Vehicle color-24 dataset

3.1 Data collection

Existing vehicle datasets have at most 13 types of vehicle color classification, and there is only one vehicle in each image. Therefore, it is difficult to be used as a benchmark dataset for vehicle color recognition or multi-target detection, which is required by practical applications. To address this challenge, we have constructed a benchmark vehicle dataset with 24 colors, namely Vehicle Color-24. The main contents of Vehicle Color-24 dataset in this paper are detailed as follows:

1) We collected 3-5 minutes video clips, totaling 100 hours, and downloaded some images from different road surveillance cameras in Xi'an with the front shooting direction. The key frames from the video and all images downloaded from surveillance services together compose into the Vehicle Color-24 dataset with 10091 images totally. Therein, each image contains 1–9 vehicles, and only one dominant color per vehicle with a resolution of 1920×1080 pixels is noticed. The Vehicle Color-24 dataset cover 24 color types, such as red, dark-red, pink, orange, dark-orange, red-orange, yellow, lemon-yellow, earthy-yellow, green, dark-green, grass-green, cyan, blue, dark-blue, purple, black, white, silver-gray, gray, dark-gray, champagne, brown and dark-brown, amounting to 31232 vehicles.

2) We have analyzed our dataset and shown the distribution in Fig. 1, and shown every class of vehicle color and the amount of corresponding vehicles in Table 1. From the Table 1, the white color class with 11827 vehicles is the head, accounting for 37.87% of the total; while the purple color

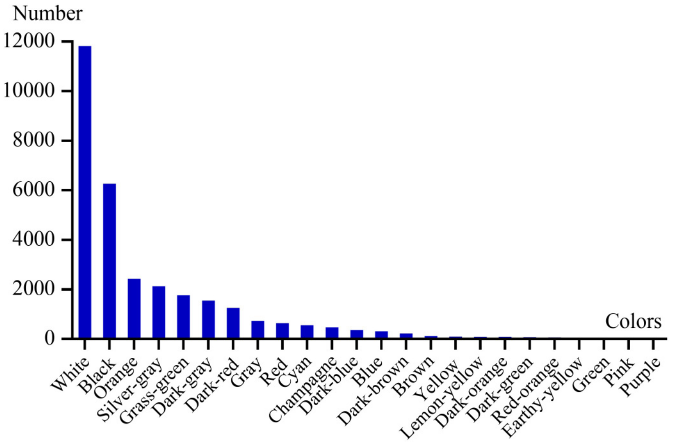


Fig. 1 24 vehicle colors distribution diagram

Table 1 Number of 24 vehicle colors

Color	Number	Percentage/%	Color	Number	Percentage/%
White	11827	37.87	Blue	316	1.01
Black	6270	20.08	Dark-brown	230	0.74
Orange	2431	7.78	Brown	118	0.38
Silver-gray	2125	6.80	Yellow	100	0.32
Grass-green	1766	5.65	Lemon-yellow	92	0.29
Dark-gray	1555	4.98	Dark-orange	90	0.29
Dark-red	1263	4.04	Dark-green	70	0.22
Gray	736	2.36	Red-orange	63	0.20
Red	644	2.06	Earthy-yellow	52	0.17
Cyan	553	1.77	Green	50	0.16
Champagne	466	1.49	Pink	35	0.11
Dark-blue	365	1.17	Purple	15	0.04

class with 15 vehicles as the tail shares of 0.04%. The distribution of the Vehicle Color-24 dataset shows long-tailed in nature.

3) In particular, all the samples of this dataset are about real traffic scenarios and cover many main types of vehicles such as sedan, truck, van and bus, etc.. When we select samples, we take into account as many scenarios as possible, such as good weather, bad weather, haze, rainy weather, low illumination, over exposure, etc., in order to keep more consistent with the real world. Two samples of every vehicle color class are shown in Table 2.

Information about the dataset is summarized as follows:

- Dataset size – 10091 images
- Test Dataset size – 1000 images
- Image size – 1920 × 1080 pixels
- Condition – Day time
- Data source – Road Surveillance Cameras

3.2 Preprocessing

Since our dataset samples were collected from real world, and then problems like haze, low illumination and overexposure in the sample images are inevitable. Therefore, we first conducted haze removal and illumination adjustment for all samples in the dataset. Our preprocessing methods of haze removing and illumination adjusting are from [37].

- Image dehaze

Haze weather tends to affect the performance of vehicle color recognition. In order to remove haze from the image, we use Eq. (1) to restore the clean image without haze.

$$J(x)=\frac{I(x)-A}{\max(t(x),t_0)}+A, \tag{1}$$

where $J(x)$ is the haze-free image to be restored, $I(x)$ is the degraded image with haze, A is the component of global atmospheric light, $t(x)$ is the transmittance, t_0 is the threshold. Our paper takes t_0=0.1. Figure 2 shows the comparison between before and after dehazing of a sample image in our dataset.

- Illumination adjustment

Illumination instability tends to affect the performance of vehicle color recognition. In order to adjust the illumination of the image, this paper makes a linear adjustment on the RGB three channels for image, as shown in Eq. (2) [37].

$$Val_i=\alpha\times Col_i+\beta, i=R,G,B, \tag{2}$$

where Col_i represents the R, G or B value of the original image respectively, Val_i represents the corresponding adjusted R, G or B value, α and β are the hyperparameters.

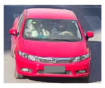
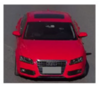
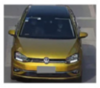
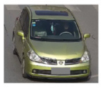
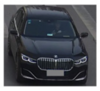
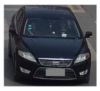
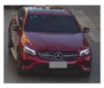
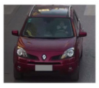
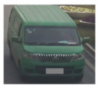
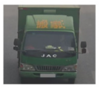
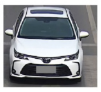
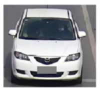
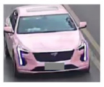
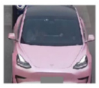
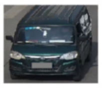
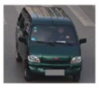
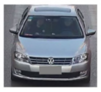
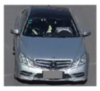
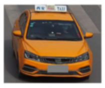
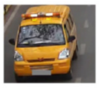
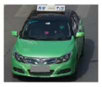
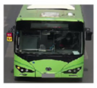
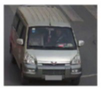
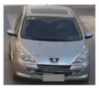
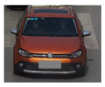
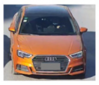
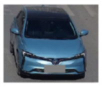
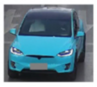
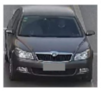
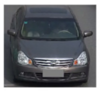
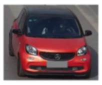
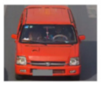
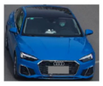
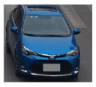
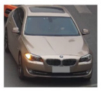
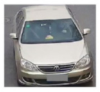
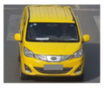
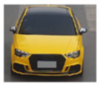
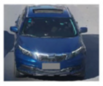
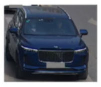
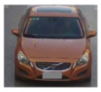
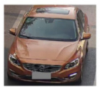
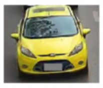
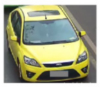
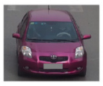
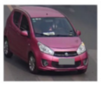
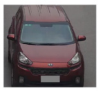
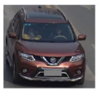
When the brightness of the image needs to be enhanced, set α= 1.5 and β= 0, so that the characteristic gap between different colors will be more obvious. When the image exposure needs to be reduced, set α= 0.8 and β= –10, the exposure of the entire image can be ameliorated. Figure 3 shows the comparison between before and after illumination adjustment in our dataset.

4 Methodology

4.1 Framework overview

To efficiently and accurately recognize the vehicle colors to

Table 2 Sample of vehicle color-24 benchmark dataset

Color	Sample picture	Color	Sample picture	Color	Sample picture
Red	 	Earthy-yellow	 	Black	 
Dark-red	 	Green	 	White	 
Pink	 	Dark-green	 	Silver-gray	 
Orange	 	Grass-green	 	Gray	 
Dark-orange	 	Cyan	 	Dark-gray	 
Red-orange	 	Blue	 	Champagne	 
Yellow	 	Dark-blue	 	Brown	 
Lemon-yellow	 	Purple	 	Dark-brown	 

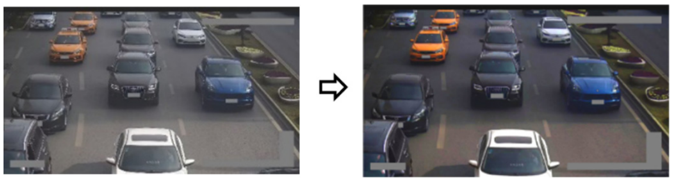


Fig. 2 Comparison before and after removal of haze

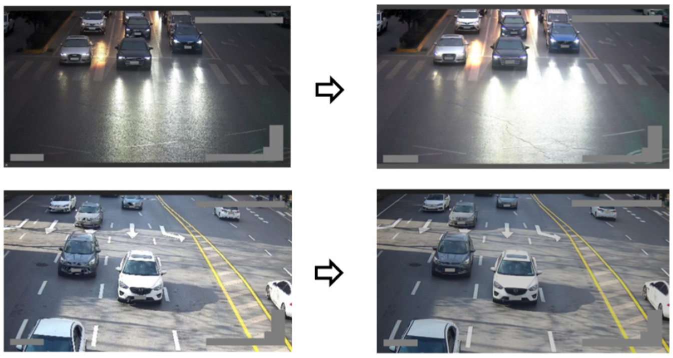


Fig. 3 Comparison before and after brightness adjustment

assist in tracking the target vehicle, in this part, we propose our SMNN-MSFF model for VCR. Intuitively, Figure 4 illustrates the framework of our SMNN-MSFF, which adopts the efficient Faster-RCNN as the backbone, especially incorporates multi-scale feature fusion into the network to obtain a high-precision recognition rate. Furthermore, we introduce a smooth modulation strategy to reduce the problem of model training degradation caused by imbalanced data categories to a certain extent. In general, our SMNN-MSFF mainly includes three parts, i.e., multi-scale feature fusion, suggesting box generation, and smooth modulation.

- **Multi-scale feature fusion** In order to simultaneously learn the geometric detail features and high-level semantic features of vehicles, we introduce multi-scale feature fusion into the backbone network to enrich a single feature map.
- **Suggesting box generation** To obtain the information of vehicle location, we first adopt a suggesting box generation module to obtain ROI from the feature maps, and then train a classifier to divide these ROIs into background and foreground. At the same time, selecting bounding boxes on the location of these ROIs is carried out by using regression.
- **Smooth modulation** We optimize the network by improving the Smooth L1 loss function, to some extent, we have paid more attention to the puzzle of class imbalance in the fine color recognition of vehicles. Smooth L1 loss is used to distinguish and optimize different types of objects, and to adjust the target frame accurately.

4.2 Multi-scale feature fusion

Previous studies [38] have shown that the shallow layers of deep neural networks are good at perceiving low-level visual features, including color, texture, etc. Naturally, model can accurately recognize the color of the vehicle by accurately perceiving the underlying visual features such as color and texture. Therefore, this paper uses as shallow CNN model, namely VCR-ResNet, as the visual encoder. In particular, our VCR-ResNet is a lightweight variant of the ResNet-50 network [39], and its structure is shown in Fig. 4, which mainly consists of 4 residual blocks with a total of 42 layers, and the detail model parameters are presented in Table 3. Besides, Figure 5 presents the feature visualization results output by each network layer of VCR-ResNet.

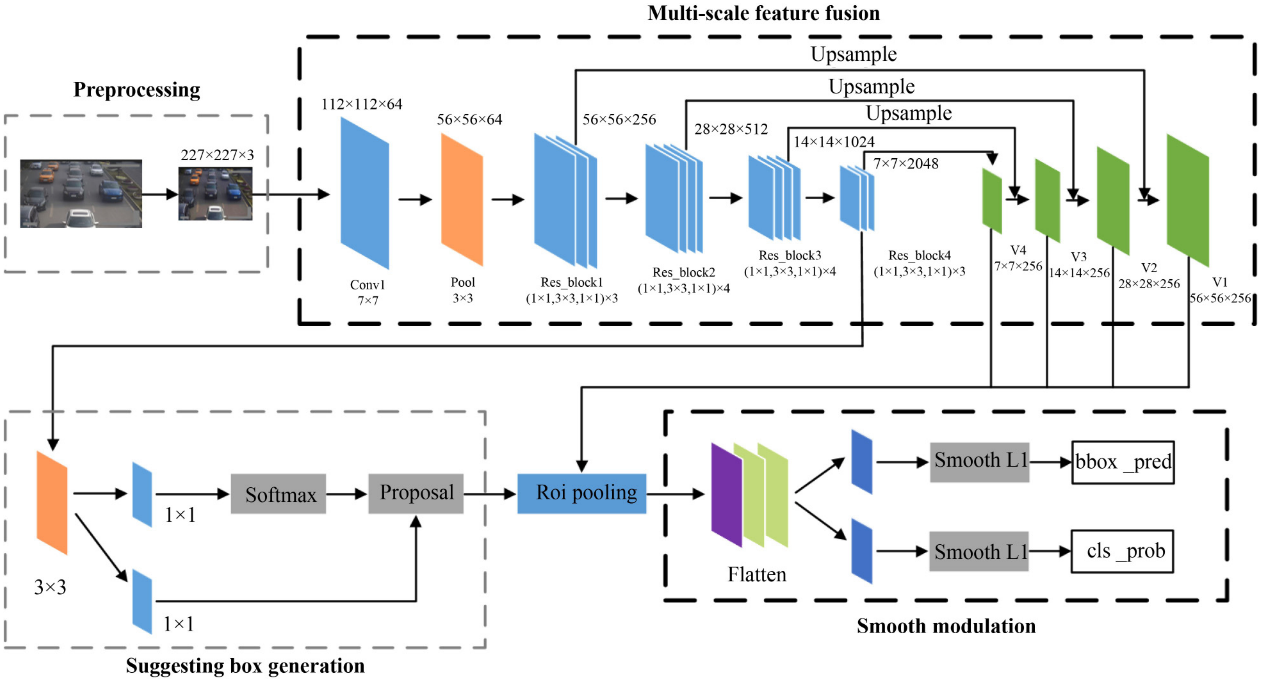


Fig. 4 The framework overview for our SMNN-MSFF

Table 3 Structural parameters of VCR-ResNet

Operation	Kernel	Stride	Feature map size	Dimension
Input	—	—	227×227×3	—
Conv1	7×7	2	112×112×64	9408
Maxpool	3×3	2	56×56×64	—
Res_block1	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times3$	1	56×56×256	442368
Res_block2	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times4$	1	28×28×512	4718592
Res_block3	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times4$	1	14×14×1024	18874368
Res_block4	$\begin{bmatrix}1\times1\\3\times3\\1\times1\end{bmatrix}\times3$	1	7×7×2048	56623104
Average pool	7×7	1	1×1×2048	—
Fc	—	—	1000	2048000

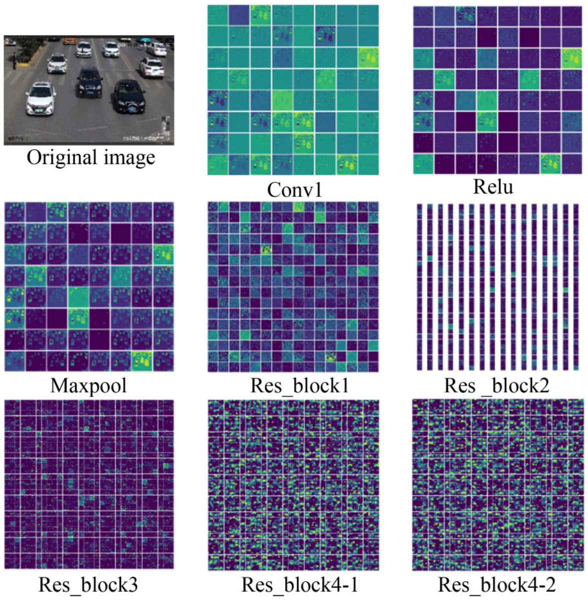

Original image
Conv1
Relu
Maxpool
Res_block1
Res _block2
Res_block3
Res_block4-1
Res_block4-2

Fig. 5 Visualization results of each layer of VCR-ResNet

In addition, to simultaneously learn and utilize low-level geometric detail features and high-level semantic features for object classification, we incorporate the multi-scale feature fusion into the backbone to enrich the single feature map. Specifically, we embed the Feature Pyramid Networks (FPN) [40] in our VCR-ResNet encoder. The FPN sequentially fuses the features output from each residual layer, they contain local detail combined features from earlier layers and high-order semantic features from the top layers. As shown in Fig. 6, the introduced FPN contains eight convolutional layers and three upsampling operations. For each upsampling operation, it utilizes the features output by each residual layer as input, aiming to map it to a high-dimensional feature space for fusion with earlier features. Formally, we assume that the output feature map of ith residual block of VCR-ResNet is $v_i^{H\times W\times C}, i\in[1,4]$, using v_i as the FPN's input, we will get the corresponding fusion feature V_i as follows:

$$V_i=\sum_{j=i}^{4} f_i(\theta_i,v_i), \tag{3}$$

where f_i means the transformation function of each fusion.

Afterwards, each output results of each layer are all input to the ROI pooling layer for the screening of regions of interest. Naturally, the softmax regression is selected as our classification function.

4.3 Smooth modulation

In general, the loss function is to calculate the difference between the real output and the ideal output. The choice of the loss function directly determines the performance of the network model. For object detection and recognition, four excellent loss functions are often used for this task, including cross entropy loss (CE Loss), L1 loss, mean square error loss (MSE Loss) and focal loss [18]. As the Smooth L1 loss function can prevent gradient explosion and has the advantages of both CE loss and L1 loss, this paper adopts the

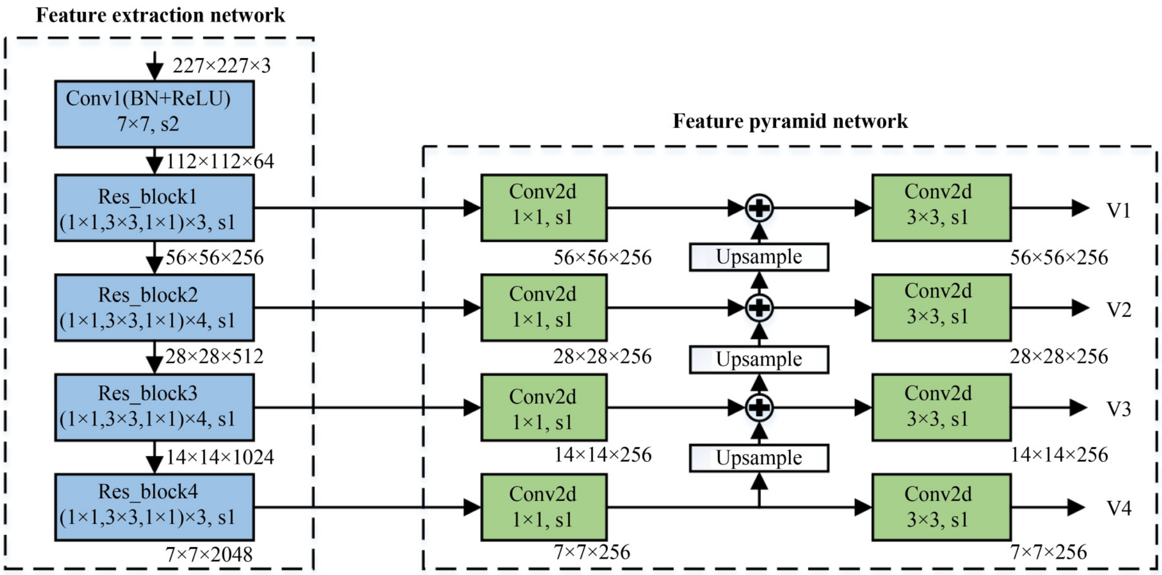


Fig. 6 The structure of multi-scale feature fusion

improved Smooth L1 loss for back propagation to update model parameters. Because, there is a long-tail distribution of samples in the VCR dataset, to alleviate this problem and inspired by the optimized strategy. We also add a adjust parameter β into the original Smooth L1 loss function, can better capture the characteristics of small-scale classes by β. The definition of the optimized Smooth L1 Loss (this is also called a VCR-Loss for differentiation purposes) is as follow,

$$\text{loss}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 0.5*(y_i - f(x_i))^2/\beta, & \text{if } |y_i - f(x_i)| < \beta, \\ |y_i - f(x_i)| - 0.5^*\beta, & \text{otherwise}, \end{cases} \tag{4}$$

where y_i is the true value, $f(x_i)$ is the predicted value or ideal output. β is the adjust parameter, which can flexibly adapt to the number of samples of different categories and cause different loss values. Empirically, the β is set as 0.11 in our experiments. Especially, our VCR-Loss has a better performance than focal loss through ablation experiments.

5 Experiments

In this section, we first introduce the experimental setup, involving the dataset details, the implementation details and the evaluation metric. Then, we performed some ablation studies to verify the rationality and effectiveness of each element of our model. Finally, we present and analyze the experimental results in detail.

5.1 Experimental setup

Datasets To fit the problem we are dealing with, we first conduct a series of experiments on our collected Vehicle Color-24 dataset, and the characteristics of this dataset are already described in detail in Section 3. Besides, three well-studied VCR datasets are also used to evaluate the recognition performance of the proposed SMNN-MSFF, including:

- The C-dataset [3], which contains 15601 images with 8 colors, involving gray (3046), red (1941), cyan (281), blue (1086), white (4742), black (3442), yellow (581) and green (482). The resolution of the samples is 1920 × 1080 pixels.
- The J-dataset [4], which contains 6819 images with 7 colors, including black (1906), gray (590), silver (950), white (2346), blue (476), red (458) and yellow (93). The resolution of the samples is 780 × 350 pixels.
- The T-dataset [5], which contains 7000 images with 13 colors, there are 2500 images in the test dataset, consisting of silver (646), white (502), black (310), gray (310), sky blue (124), red (122), blue (118), brown (74), orange (72), pink (62), green (58), yellow (55) and purple (47). The resolution of the samples is 1920 × 1080 pixels.

Implementation details We implement our models on the Pytorch, and the experimental platform is set as follows: the operating system is Windows 10 with Intel Core i7-8700K, the memory is 16 GB, the GPU is NVIDIA GTX 1070Ti, and the video memory is 8 GB.

To optimize and validate the proposed model, we divide the Vehicle Color-24 dataset into training set, verification set and test set at a ratio of 8:1:1. In addition, the initial learning rate is set to 0.0001 in the training phase, and the batch size is 4, the epoch is 50 and the confidence threshold is 0.5.

Evaluation metric Following the previous studies, mean average precision (mAP) is also used to measure the performance of the network for vehicle color recognition. AP value is calculated by the area under the P(R) curve formed by Precision and Recall under different thresholds through interpolation solution, as shown in equation 4.

$$AP = \int_0^1 P(R)\mathrm{d}R = \sum_{k=1}^{N} \max_{k_1 \geqslant k} P(k_1)\Delta R(k), \tag{5}$$

where $\max_{k_1 \geqslant k} P(k_1)$ represents the maximum precision rate of all recall rates which are not less than a certain recall rate R, and $\Delta R(k)$ represents the change value of recall rate. The mAP is the average of the AP values of all classes.

5.2 Ablation experiments

The effect of VCR-ResNet In order to evaluate the effect of our proposed visual encoder, i.e., VCR-ResNet, we compared the recognition performance by Faster-RCNN with 6 different visual encoders, including VGG12, VGG16, ResNet34, ResNet50, MobileNet and our VCR-ResNet. As shown in Table 4, it's obvious that the VCR-ResNet adopted in this paper achieved the best performance with an average detection accuracy of 68.17%, which is a relative increase of 9.58%. It means that the introduced VCR-ResNet could improve the recognition performance.

The effect of MSFF In order to demonstrate the effectiveness of our proposed multi-scale feature fusion (MSFF), we compared the method with using MSFF with the method does not using MSFF. As shown in Table 5, when we use the MSFF, the average detection accuracy reaches 74.38%, which achieves a relative increase of 15.79%. It indicates that the proposed MSFF is effect to VCR.

The effect of VCR-Loss We conducted ablation experiments to verify the influence of our VCR-Loss by comparing the performance with using different classic loss functions, including CE Loss, L1 Loss, MSE Loss and Focal Loss on the basis of multi-scale feature fusion. It can be seen from Table 5 that the VCR-Loss function in this paper could achieve the best performance with an average detection accuracy of 94.96%, which is a relative increase of 36.37%.

In addition, the comparison of training loss values with different models is shown in Fig. 7. It can be seen that the training loss values of the model decreased significantly after using the VCR-ResNet, indicating that the model performance has been improved.

5.3 Comparative experiments and analysis

To better evaluate our Vehicle Color-24 dataset, and comprehensively verify the capability of the proposed SMNN-MSFF, we conducted a series of comparative recognition experiments on different VCR datasets, including C-dataset, J-dataset, T-dataset, and our Vehicle Color-24 dataset.

- 24-Color recognition results

In this subsection, we evaluated the proposed Vehicle Color-24 dataset by using several excellent target detection model,

Table 4 Comparison table of recognition accuracy of feature extraction network ablation experiments

	Faster-RCNN	VGG12	VGG16	ResNet34	ResNet50	MobileNet	+VCR-ResNet
White	0.84	0.64	0.83	0.81	0.82	0.8	0.85
Black	0.82	0.6	0.83	0.79	0.81	0.74	0.84
Orange	0.81	0.75	0.83	0.8	0.81	0.79	0.81
Silver-gray	0.77	0.62	0.78	0.7	0.71	0.71	0.73
Grass-green	0.70	0.72	0.76	0.75	0.78	0.77	0.8
Dark-gray	0.66	0.57	0.68	0.63	0.66	0.62	0.67
Dark-red	0.78	0.61	0.76	0.73	0.72	0.70	0.72
Gray	0.18	0.50	0.26	0.53	0.55	0.58	0.65
Red	0.60	0.58	0.60	0.70	0.71	0.70	0.73
Cyan	0.75	0.72	0.80	0.77	0.78	0.81	0.80
Champagne	0.63	0.59	0.76	0.72	0.7	0.67	0.70
Dark-blue	0.66	0.63	0.75	0.63	0.68	0.62	0.70
Blue	0.73	0.62	0.77	0.65	0.65	0.61	0.66
Dark-brown	0.45	0.48	0.29	0.61	0.63	0.63	0.69
Brown	0.30	0.52	0.07	0.57	0.56	0.66	0.68
Yellow	0.51	0.70	0.56	0.51	0.49	0.49	0.44
Lemon-yellow	0.87	0.70	0.92	0.55	0.57	0.56	0.59
Dark-orange	0.65	0.64	0.52	0.58	0.57	0.51	0.62
Dark-green	0.38	0.63	0.75	0.54	0.53	0.54	0.57
Red-orange	0.24	0.58	0.43	0.49	0.51	0.5	0.51
Earthy-yellow	0.62	0.69	0.33	0.52	0.51	0.55	0.6
Green	0.61	0.73	0.97	0.63	0.65	0.67	0.69
Pink	0.50	0.68	0.50	0.70	0.72	0.73	0.71
Purple	0	0.47	0	0.55	0.57	0.58	0.60
mAP	**58.59%**	**62.38%**	**61.49%**	**64.42%**	**65.38%**	**64.75%**	**68.17%**
Relative increase	**0%**	**3.79%**	**2.90%**	**5.83%**	**6.79%**	**6.16%**	**9.58%**

Table 5 Comparison of recognition accuracy between multi-scale feature fusion and loss function ablation experiments

	Faster-RCNN	+FPN	L1 Loss	CE Loss	MSE Loss	Focal Loss	+VCR-Loss
White	0.84	0.96	0.96	0.95	0.99	0.98	0.98
Black	0.82	0.94	0.94	0.90	0.98	0.97	0.97
Orange	0.81	0.88	0.88	0.87	0.97	0.98	0.98
Silver-gray	0.77	0.85	0.85	0.86	0.93	0.97	0.96
Grass-green	0.70	0.81	0.81	0.83	0.95	0.98	0.98
Dark-gray	0.66	0.73	0.73	0.78	0.88	0.94	0.94
Dark-red	0.78	0.83	0.83	0.82	0.87	0.97	0.98
Gray	0.18	0.69	0.69	0.76	0.79	0.82	0.89
Red	0.60	0.78	0.78	0.75	0.74	0.96	0.96
Cyan	0.75	0.81	0.81	0.83	0.84	0.98	0.97
Champagne	0.63	0.76	0.76	0.79	0.8	0.94	0.91
Dark-blue	0.66	0.77	0.77	0.74	0.75	0.97	0.96
Blue	0.73	0.77	0.77	0.78	0.79	0.97	0.97
Dark-brown	0.45	0.79	0.79	0.77	0.77	0.88	0.97
Brown	0.30	0.71	0.71	0.75	0.78	0.80	0.88
Yellow	0.51	0.30	0.30	0.79	0.81	0.95	0.97
Lemon-yellow	0.87	0.77	0.77	0.85	0.87	0.99	0.99
Dark-orange	0.65	0.70	0.70	0.81	0.48	0.94	0.96
Dark-green	0.38	0.71	0.71	0.8	0.89	0.91	0.94
Red-orange	0.24	0.54	0.54	0.71	0.57	0.94	0.99
Earthy-yellow	0.62	0.71	0.71	0.73	0.78	0.92	0.97
Green	0.61	0.73	0.73	0.75	0.79	0.89	0.93
Pink	0.50	0.66	0.66	0.63	0.55	0.90	0.94
Purple	0	0.65	0.65	0.5	0.57	0.48	0.80
mAP	**58.59%**	**74.38%**	**74.38%**	**78.13%**	**79.75%**	**91.79%**	**94.96%**
Relative increase	**0%**	**15.79%**	**15.79%**	**19.54%**	**21.16%**	**33.20%**	**36.37%**

including Faster-RCNN [12], SSD [13], YOLO-v3 [14], YOLO-v4 [15], Efficient-Det [16], Center-Net [17], and Retina-Net [18]. In addition, we also compared the recognition performance of our SMNN-MSFF with these methods.

Table 6 presents the recognition results. It can be seen that the proposed SMNN-MSFF performs best in this task. The average recognition accuracy of our method is 94.96%, which is 33.47% higher than Faster-RCNN, 16.83% higher than the

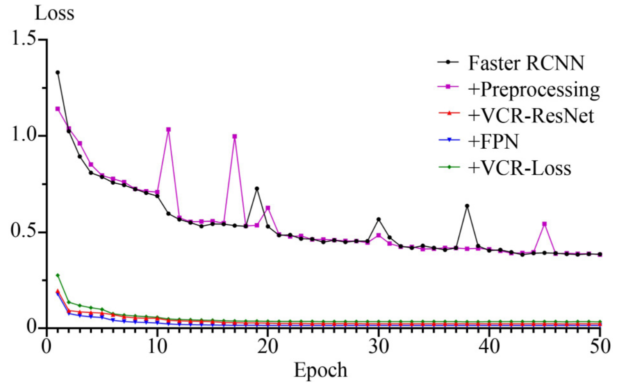


Fig. 7 Comparison of training loss of different models

SSD method, 18.58% higher than the YOLO-v3 method, 32.19% higher than the YOLO-v4 method, 34.10% higher than the Efficient-Det method, 33.89% higher than the Center-Net method, and 3.27% higher than the Retina-Net method.

Furthermore, Figure 8 illustrated the real-time recognition results of our SMNN-MSFF. It can be seen that the recognition accuracy is higher when recognizing categories with more samples of white and black, while the recognition accuracy is relatively low when recognizing categories with fewer samples of pink and dark-orange.

● 8-Color recognition results

To further verify the performance of our proposed SMNN-MSFF, we also conducted a 8-Color recognition experiment on the most well-studied benchmark dataset, i.e., C-dataset. To better verify the SMNN-MSFF's effectiveness, we compared the recognition results with several recent state-of-the-art methods on the current 8-Color vehicle recognition, including the algorithm proposed by Chen et al. [3], Hu et al. [7], Fu et al. [10].

As shown in Table 7, we can easily find that our methods perform best in each category of 8-Color. In detail, the average recognition accuracy of our method is 97.25%, which is 4.62% higher than the algorithm proposed by Chen et al., 2.87% higher than the algorithm proposed by Hu et al., and 0.12% higher than the algorithm proposed by Fu et al.. Compared with a few multi-sample classes, although our recognition rate is slightly lower, the average recognition performance of our method is significantly better than state-of-the-art methods.

● Comparison of different VCR datasets

To verify the generalization performance of the proposed method, we also conducted a series of comparative experiments on different datasets with different color categories, i.e., C-dataset with 8 colors, J-dataset with 7 colors, T-dataset with 13 colors. Figure 9 presents the mAP of SMNN-MSFF in the above datasets. The average detection accuracy on the C-dataset is 92.63%. The average detection accuracy on the J-dataset is 97.85%. The average detection accuracy on the T-dataset is 90.62%. It can be seen that the average recognition accuracy of our model has the best performance.

● Real-time performance analysis

For further analysis the real-time performance of the proposed method, we also compared the recognition speed with other methods using the CPU device. The input pixel of the image is 1920 × 1080, and the comparison results are shown in Table 8. The recognition speed of our model on the CPU device is 1.021 seconds, which is lower than other methods. It can conclude that our method has relatively better real-time performance.

Table 6 Comparison of recognition accuracy of 24 colors in different network classifications

	Faster-RCNN [12]	SSD [13]	YOLO-v3 [14]	YOLO-v4 [15]	Efficient-Det [16]	Center-Net [17]	Retina-Net [18]	**SMNN-MSFF**
White	0.84	0.96	0.97	0.98	0.95	0.97	0.98	**0.98**
Black	0.82	0.95	0.96	0.96	0.93	0.94	0.97	**0.97**
Orange	0.81	0.96	0.97	0.98	0.96	0.97	0.98	**0.98**
Silver-gray	0.77	0.91	0.92	0.92	0.87	0.88	**0.97**	0.96
Grass-green	0.70	0.96	0.97	0.98	0.95	0.97	0.98	**0.98**
Dark-gray	0.66	0.84	0.82	0.93	0.73	0.73	0.94	**0.94**
Dark-red	0.78	0.93	0.94	0.94	0.91	0.92	0.97	**0.98**
Gray	0.18	0.54	0.50	0.40	0.28	0.28	0.82	**0.89**
Red	0.60	0.88	0.88	0.81	0.79	0.77	0.96	**0.96**
Cyan	0.75	0.92	0.93	0.92	0.82	0.89	**0.98**	0.97
Champagne	0.63	0.81	0.83	0.77	0.66	0.76	**0.94**	0.91
Dark-blue	0.66	0.86	0.85	0.87	0.76	0.77	**0.97**	0.96
Blue	0.73	0.87	0.85	0.87	0.75	0.82	0.97	**0.97**
Dark-brown	0.45	0.71	0.68	0.60	0.49	0.64	0.88	**0.97**
Brown	0.30	0.58	0.52	0.25	0.36	0.25	0.80	**0.88**
Yellow	0.51	0.79	0.72	0.56	0.47	0.42	0.95	**0.97**
Lemon-yellow	0.87	0.93	0.84	0.77	0.66	0.64	0.99	**0.99**
Dark-orange	0.65	0.78	0.66	0.60	0.66	0.53	0.94	**0.96**
Dark-green	0.38	0.58	0.63	0.00	0.23	0.66	0.91	**0.94**
Red-orange	0.24	0.61	0.61	0.05	0.33	0.13	0.94	**0.99**
Earthy-yellow	0.62	0.74	0.69	0.43	0.43	0.29	0.92	**0.97**
Green	0.61	0.74	0.77	0.54	0.33	0.40	0.89	**0.93**
Pink	0.50	0.71	0.75	0.03	0.18	0.01	0.90	**0.94**
Purple	0.00	0.19	0.08	0.00	0.10	0.00	0.48	**0.80**
mAP	**58.59%**	**78.13%**	**76.38%**	**62.77%**	**60.86%**	**61.07%**	**91.79%**	**94.96%**

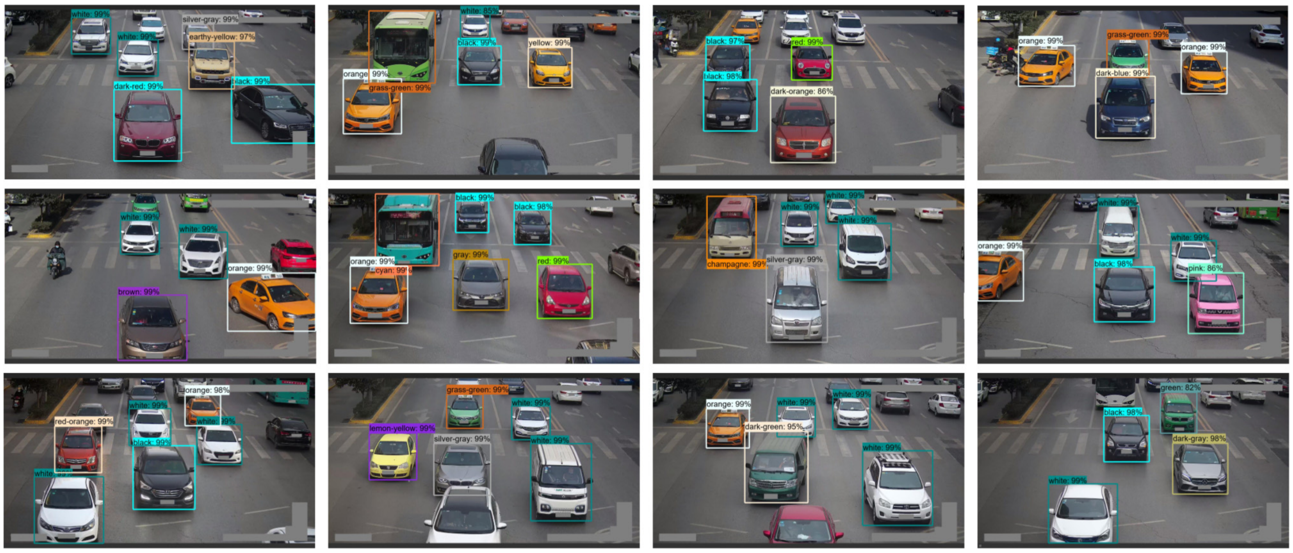


Fig. 8 Vehicle recognition results with 24 colors

Table 7 Comparison of recognition accuracy of 8 colors in different networks

	Chen et al. [3]	Hu et al. [7]	Fu et al. [10]	**SMNN-MSFF**
White	0.94	0.96	0.98	0.99
Black	0.97	0.97	0.99	0.98
Gray	0.85	0.87	0.96	0.96
Red	0.99	0.99	0.99	0.98
Cyan	0.98	0.99	0.96	0.98
Blue	0.95	0.97	0.94	0.97
Yellow	0.95	0.97	0.98	0.96
Green	0.78	0.83	0.97	0.96
mAP	**92.63%**	**94.38%**	**97.13%**	**97.25%**

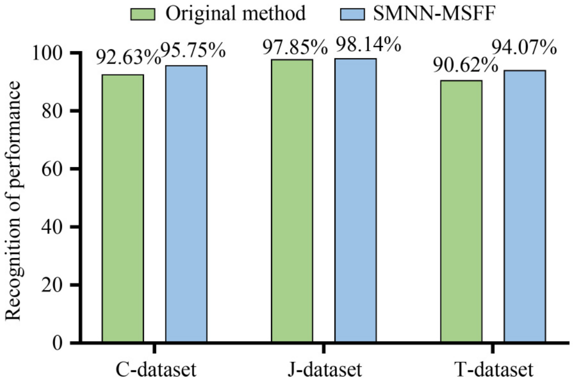


Fig. 9 Recognition of performance on different datasets

Table 8 Comparison of different network recognition speeds (using the CPU device)

Algorithm	Faster-RCNN [12]	SSD [13]	YOLO-v3 [14]	YOLO-v4 [15]	Efficient-Det [16]	Center-Net [17]	Retina-Net [18]	**SMNN-MSFF**
Inference time/s	2.670	1.689	1.265	0.989	1.062	1.233	1.359	1.021

6 Conclusion

In this paper, for studying vehicle color recognition, we have built a new vehicle color benchmark dataset which is called Vehicle Color-24, with 10091 images with 24 color types. And the Vehicle Color-24 dataset covers most classes till now. For addressing long-tail recognition, we propose a method of SMNN-MSFF with three network modules: feature extraction network, the suggesting frame generation, and the adjustment of the loss function. The experimental results show that the mAP of our method in our paper is 94.96% in recognizing 24 types of colors and the mAP of recognizing 8 types of colors is 97.25%. We have further improved the average detection accuracy, and better meet the requirements for fine classification of vehicle colors. However, since the actual environment can be affected by unpredictable factors and the long tail effect exists in vehicle color distribution, further efforts to improve the fine recognition of vehicle color are still required. And we will continue studying the solution of class imbalance, because the vehicle color is diverse, the vehicle color dataset must have the characteristics of the long-tail distribution.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 62071378), the Shaanxi Province International Science and Technology Cooperation Program (2022KW-04), and the Xi'an Science and Technology Plan Project (21XJZZ0072).

References

1. Ke X, Zhang Y F. Fine-grained vehicle type detection and recognition based on dense attention network. Neurocomputing, 2020, 399: 247–257
2. Tariq A, Khan M Z, Khan M U G. Real time vehicle detection and colour recognition using tuned features of faster-RCNN. In: Proceedings of the 1st International Conference on Artificial Intelligence and Data Analytics. 2021, 262–267
3. Chen P, Bai X, Liu W Y. Vehicle color recognition on urban road by feature context. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(5): 2340–2346
4. Jeong Y, Park K H, Park D. Homogeneity patch search method for

voting-based efficient vehicle color classification using front-of-vehicle image. Multimedia Tools and Applications, 2019, 78(20): 28633–28648

5. Tilakaratna D S B, Watchareeruetai U, Siddhichai S, Natcharapinchai N. Image analysis algorithms for vehicle color recognition. In: Proceedings of 2017 International Electrical Engineering Congress. 2017, 1–4
6. Dule E, Gökmen M, Beratoğlu M S. A convenient feature vector construction for vehicle color recognition. In: Proceedings of the 11th WSEAS International Conference on Nural Networks and 11th WSEAS International Conference on Evolutionary Computing and 11th WSEAS International Conference on Fuzzy Systems. 2010, 250–255
7. Hu C P, Bai X, Qi L, Chen P, Xue G J, Mei L. Vehicle color recognition with spatial pyramid deep learning. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(5): 2925–2934
8. Rachmadi R F, Purnama I K E. Vehicle color recognition using convolutional neural network. 2015, arXiv preprint arXiv: 1510.07391
9. Zhuo L, Zhang Q, Li J F, Zhang J, Li X G, Zhang H. High-accuracy vehicle color recognition using hierarchical fine-tuning strategy for urban surveillance videos. Journal of Electronic Imaging, 2018, 27(5): 051203
10. Fu H Y, Ma H D, Wang G Y, Zhang X M, Zhang Y F. MCFF-CNN: multiscale comprehensive feature fusion convolutional neural network for vehicle color recognition based on residual learning. Neurocomputing, 2020, 395: 178–187
11. Nafzi M, Brauckmann M, Glasmachers T. Vehicle shape and color classification using convolutional neural network. 2019, arXiv preprint arXiv: 1905.08612
12. Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149
13. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: single shot MultiBox detector. In: Proceedings of the 14th European Conference on Computer Vision. 2016, 21–37
14. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016, 779–788
15. Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection. 2020, arXiv preprint arXiv: 2004.10934
16. Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 10778–10787
17. Zhou X Y, Wang D Q, Krähenbühl P. Objects as points. 2019, arXiv preprint arXiv: 1904.07850
18. Lin T Y, Goyal P, Girshick R, He K M, Dollár P. Focal loss for dense object detection. In: Proceedings of 2017 IEEE International Conference on Computer Vision. 2017, 2999–3007
19. Tang K H, Huang J Q, Zhang H W. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: Proceedings of the 34th Conference on Neural Information Processing Systems. 2020
20. Wang X G, Bai X, Liu W Y, Latecki L J. Feature context for image classification and object detection. In: Proceedings of the CVPR 2011. 2011, 961–968
21. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems. 2012, 1106–1114
22. Cui Y, Jia M L, Lin T Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 9260–9269
23. Cao K D, Wei C L, Gaidon A, Arechiga N, Ma T Y. Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 140
24. Ren M Y, Zeng W Y, Yang B, Urtasun R. Learning to reweight examples for robust deep learning. In: Proceedings of the 35th International Conference on Machine Learning. 2018
25. Shu J, Xie Q, Yi L X, Zhao Q, Zhou S P, Xu Z B, Meng D Y. Meta-weight-net: learning an explicit mapping for sample weighting. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 172
26. Jamal M A, Brown M, Yang M H, Wang L Q, Gong B Q. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 7607–7616
27. Kang B Y, Xie S N, Rohrbach M, Yan Z C, Gordo A, Feng J S, Kalantidis Y. Decoupling representation and classifier for long-tailed recognition. In: Proceedings of the 8th International Conference on Learning Representations. 2020
28. Zhou B Y, Cui Q, Wei X S, Chen Z M. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 9716–9725
29. Yin X, Yu X, Sohn K, Liu X M, Chandraker M. Feature transfer learning for face recognition with under-represented data. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 5697–5706
30. Liu J L, Sun Y F, Han C C, Dou Z P, Li W H. Deep representation learning on long-tailed data: a learnable embedding augmentation perspective. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 2967–2976
31. Liu Z W, Miao Z Q, Zhan X H, Wang J Y, Gong B Q, Yu S X. Large-scale long-tailed recognition in an open world. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 2532–2541
32. Chu P, Bian X, Liu S P, Ling H B. Feature space augmentation for long-tailed data. In: Proceedings of the 16th European Conference on Computer Vision. 2020, 694–710
33. Menon A K, Jayasumana S, Rawat A S, Jain H, Veit A, Kumar S. Long-tail learning via logit adjustment. In: Proceedings of the International Conference on Learning Representations. 2020
34. Xiang L Y, Ding G G, Han J G. Learning from multiple experts: self-paced knowledge distillation for long-tailed classification. In: Proceedings of the 16th European Conference on Computer Vision. 2020, 247–263
35. Li Y, Wang T, Kang B Y, Tang S, Wang C F, Li J T, Feng J S. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 10988–10997
36. Wang X D, Lian L, Miao Z Q, Liu Z W, Yu S X. Long-tailed recognition by routing diverse distribution-aware experts. 2021, arXiv preprint arXiv: 2010.01809
37. Xue X Q, Ding J K, Shi Y J. Research and application of illumination processing method in vehicle color recognition. In: Proceedings of the 3rd IEEE International Conference on Computer and Communications. 2017, 1662–1666
38. Seifert C, Aamir A, Balagopalan A, Jain D, Sharma A, Grottel S, Gumhold S. Visualizations of deep neural networks in computer vision: a survey. In: Cerquitelli T, Quercia D, Pasquale F, eds. Transparent

Data Mining for Big and Small Data. Cham: Springer, 2017, 123–144

39. He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016, 770–778
40. Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017, 936–944

Mingdi Hu, doctor, associate professor. She obtained doctor of science degree from school of mathematics and statistics at Shaanxi Normal University, China. Her research interests include image recognition, target retrieval and classification, data enhancement, machine learning, artificial intelligence and fuzzy information processing.

Long Bai, master student of Xi’an University of Posts and Telecommunications, China. He received the BS degree in communication engineering from Xi’an University of Science and Technology, China. His research interests include machine learning, deep neural network, image target recognition and artificial intelligence.

Jiulun Fan, doctor, professor. He graduated from Xidian University, China, majoring in signal and information processing, and obtained doctor degree in engineering. His research interests include pattern recognition and image processing, fuzzy information processing theory and application, image security technology.

Sirui Zhao, doctor student of University of Science and Technology of China, China. His research interests include human-computer interaction, affective computing, computer vision and knowledge representation. He has published several papers in refereed conferences and journals, such as ACM MM2021, Neural Networks.

Enhong Chen, doctor, professor of University of Science and Technology of China, China. He is CCF Fellow, IEEE Senior Member. His research interests includes data mining and machine learning, especially social network analysis and recommender systems. He has published more than 200 papers in refereed conferences and journals, such as TKED, KDD, ICDM, NIPS.