Research on Vehicle Appearance Component Recognition Based on Mask R-CNN

Zhu Qianqian1, Liu Sen2 and Guo Weiming1

1China Automotive Technology & Research Center Co,Ltd., Tianjin, China
2Alnovation AI Empowering Business, Beijing, China

E-mail: zhuqianqian91@sina.com

Abstract. Recognition of vehicle exterior components is one of the most important core algorithms in the process of intelligent propulsion. This paper focuses on the recognition algorithm of vehicle appearance parts, which can detect the position of vehicle appearance parts and recognize the name of the parts in the image. This paper applies enterprise self-built datasets. Firstly, the vehicle close-range dataset is produced by tailoring in data enhancement. Secondly, the models based on ResNeXt-50+FPN, ResNeXt-101+FPN, ResNeXt-50 backbone network and Mask R-CNN are applied in three stages: panoramic dataset, panoramic dataset and panoramic close-range integrated dataset. Finally, the network model based on ResNeXt-101+FPN and Mask R-CNN is the best on the comprehensive dataset.

1. Introduction

In recent years, with the sustained and rapid development of China's economy and society, the number of motor vehicles in China has been growing rapidly. At the same time, people pay more and more attention to vehicle safety. Among them, the status of vehicle appearance components, such as scraping, deformation, etc. are important applications in vehicle appearance quality inspection, fixed-loss claims of automobile insurance, second-hand vehicle valuation and other business scenarios. Therefore, the recognition of vehicle appearance components is particularly important as the premise of vehicle appearance condition detection to determine the location of damage.

With the rapid development of deep learning, object detection method based on deep convolution neural network has been widely used. At present, the problem of vehicle recognition is mostly focused on the detection of vehicle, vehicle type and license plate, and the research on image recognition of vehicle appearance parts is less. Aiming at the problem of vehicle appearance component recognition, in order to avoid the overlap of location areas after vehicle appearance component detection, which will bring error to subsequent application research such as damage detection, this paper proposes an example segmentation method for components, that is, the method of image appearance component is recognized by pixel-level object segmentation.

This paper builds recognition model based on Mask R-CNN network structure. ResNeXt-50+FPN, ResNeXt-101+FPN and ResNeXt-50 are used as feature extraction networks respectively. The objective of vehicle appearance components is divided into 31 categories, including front bumper, front fender and central network. In the training phase, the self-built dataset CATARC-Vehicle of vehicle appearance components is used, and the data enhancement and reasonable adjustment of parameters are carried out to complete the target detection research in complex scenes.
2. Establishment of Recognition Model for Vehicle Appearance Components

2.1. Introduction of Mask R-CNN

Mask R-CNN network model [1] was proposed by He Kaiming in the ICCV paper. It can effectively detect targets and output high-quality instance segmentation at the same time. Mask R-CNN network model improves the network structure on Faster R-CNN network model [2], increases the output of the target mask, and enables the Mask R-CNN network model to accomplish three tasks of target recognition, candidate box regression and segmentation mask at the same time.

Mask R-CNN network model first extracts features from input images, and then inputs image features into RPN network [3] to get candidate frames. Mask R-CNN network model uses ROI Align layer to correct ROI pixels, and then uses head structure network to classify targets and regress candidate boxes. In parallel with the implementation of classification and regression problems, the branch of Mask prediction is added, and each pixel in ROI can be identified as belonging to the target category. The loss function of Mask R-CNN network consists of three parts: classification error, regression error and segmentation error. The Mask R-CNN network structure is shown in Figure 1.

![Figure 1. Mask R-CNN network structure schematic diagram](image)

2.2. Vehicle Appearance Component Recognition Model

ResNeXt [4] solves the problem of deep network degradation compared with LeNet [5], AlexNet [6], and has better results than ResNet [7] with the same number of parameters. The feature pyramid network FPN [8], which consists of special diagnostic images of different scales, retains the location information better in the low-level feature images, and can analyze the semantic information better in the high-level feature images. Combining the high-level feature images of each layer with the low-level feature images, it can better deal with small-scale targets. In addition, it can better analyze location information and semantic information.

Aiming at the problem of vehicle appearance component recognition, it needs high accuracy of model recognition and fast prediction speed of model. In order to make Mask R-CNN better applied to vehicle appearance parts target detection, this paper selected ResNeXt-50+FPN, ResNeXt-101+FPN, ResNeXt-50 backbone network as feature extraction network and combined with Mask R-CNN, and built the network structure for experiments.

3. The Dataset

3.1. Panoramic Dataset

There is no public datasets for vehicle appearance component recognition. Therefore, the dataset used in this experiment comes from the dataset CATARC-Vehicle built by the unit. In order to realize the pixel-level target detection of the vehicle appearance parts, the CATARC-Vehicle dataset adopts the form of polygon segmentation and annotation, totaling 10,026 panoramic images. Vehicle information in the image is very special, such as no serious damage, no component opening, etc. In addition, the image includes daytime and night shooting, which is very suitable for vehicle appearance detection in complex scenes. The annotation example is shown in Figure 2.
3.1.1. Description of Vehicle Type. Vehicle types in the image include car, SUV, MPV and VAN (minibus). Image data are evenly distributed in each dimension. An example diagram of vehicle type is shown in Figure 3.

3.1.2. Description of Vehicle Orientation. Vehicle shooting angle in the image includes five azimuth types: front, anterolateral, posterolateral, side and rear. Picture data are evenly distributed in all dimensions. An example diagram of vehicle angle is shown in Figure 4.

3.1.3. Type Description of Vehicle Appearance Parts. Vehicle exterior components are complex and diverse. 31 items of damage-prone exterior components are selected in the dataset. The front door and the rear door are regarded as two parts, while the left front door and the right front door are regarded as the same parts. The list of parts is shown in Table 1.
### Table 1. Vehicle appearance component list.

| Numble | Component Name    | Numble | Component Name    | Numble | Component Name    |
|--------|-------------------|--------|-------------------|--------|-------------------|
| 1      | Front Bumper      | 4      | Hood              | 7      | Rear Fender       |
| 2      | Rear Bumper       | 5      | Roof Panel        | 8      | Door Glass        |
| 3      | Intermediate Network | 6  | Front Fender      | 9      | Tires             |

#### 3.2. Close-Range Dataset

In the application scenario of vehicle appearance recognition, in order to shoot the vehicle damage more clearly, the vehicle close-range image will be taken. Therefore, the recognition model of vehicle appearance components needs to have good recognition ability for local appearance components in close-range images. Data enhancement is a commonly used method when the number of data is insufficient, such as image flipping, rotation, scaling, clipping and so on. In this paper, in order to save the cost of image annotation and achieve the goal of constructing the close-range dataset, the vehicle panoramic dataset is cut vertically and horizontally to get four close-range images, and the close-range dataset of vehicle appearance component recognition problem is produced.

The specific methods are as follows: (1) Firstly, the trained vehicle detection model is applied to identify the position of the vehicle in the picture, and the transverse and longitudinal axes are obtained by taking the vehicle center as the segmentation point; (2) the original picture is cut into four images according to the transverse and longitudinal axes; (3) in the process of cutting the corresponding labeling information, four labeling documents are updated. Among them, we should pay attention to the complex concave polygon.

The number of close-range dataset is larger than that of panoramic dataset because the actual target tends to be close-range. After data cleaning and excluding invalid pictures, 20,287 pictures were finally obtained. The clipped image is shown in Figure 5.

![Figure 5. Examples of close-range dataset](image)

#### 4. Experiments and Results Analysis

The experimental computer in this paper is configured as a high-performance computer, six NVIDIA GFORCE GTX, Centos operating system, CUDA10.1, cuDNN7.5.0, based on deep learning framework caffe2, programming environment based on Python 3.7.

Close-range images were obtained by data enhancement. There are 30,313 datas available. The datas were randomly divided into training set, verification set and test set, and the corresponding proportion was 70%, 15% and 15%. After dataset partitioning, statistics show that there are 6,951 training sets, 1,513 validation sets and 1,562 test sets in panoramic images, 14,174 training sets, 3,119 validation sets and 2,994 test sets in close-range images.

The experiment is based on dataset CATARC-Vehicle, which is divided into three stages: (1) panoramic dataset; (2) close-range dataset; (3) panoramic and close-range integrated dataset, which are
trained separately. The experimental models of each stage are based on ResNeXt-50+FPN, ResNeXt-101+FPN, ResNeXt-50 and Mask R-CNN respectively.

In order to explore a faster and concise method, experiments are carried out on three datasets, and in order to meet the needs of practical application, comprehensive datasets are used to test, which can better evaluate whether the model meets the needs of application scenarios.

4.1. First Stage

In the first stage, experiments are carried out on panoramic dataset based on the networks of ResNeXt-50+FPN, ResNeXt-101+FPN and ResNeXt-50 backbone combined with Mask R-CNN. The experimental results of target segmentation are shown in Table 2.

| backbone          | mask AP | mask AP50 | mask AP75 | mask APs | mask APm | mask APl |
|-------------------|---------|-----------|-----------|----------|----------|----------|
| Mask R-CNN        | ResNeXt-50-FPN | 0.52     | 0.67      | 0.59     | 0.15     | 0.38     | 0.63     |
| Mask R-CNN        | ResNeXt-101-FPN | 0.57     | 0.71      | 0.63     | 0.20     | 0.43     | 0.67     |
| Mask R-CNN        | ResNeXt-50     | 0.47     | 0.63      | 0.53     | 0.05     | 0.33     | 0.58     |

4.2. Second Stage

In the second stage, experiments are carried out on close-range dataset based on three backbone networks of ResNeXt-50+FPN, ResNeXt-101+FPN and ResNeXt-50 combined with Mask R-CNN. The experimental results of target segmentation are shown in Table 3.

| backbone          | mask AP | mask AP50 | mask AP75 | mask APs | mask APm | mask APl |
|-------------------|---------|-----------|-----------|----------|----------|----------|
| Mask R-CNN        | ResNeXt-50-FPN | 0.62     | 0.81      | 0.68     | 0.39     | 0.57     | 0.65     |
| Mask R-CNN        | ResNeXt-101-FPN | 0.68     | 0.86      | 0.74     | 0.51     | 0.63     | 0.71     |
| Mask R-CNN        | ResNeXt-50     | 0.56     | 0.76      | 0.61     | 0.35     | 0.51     | 0.59     |

4.3. Third Stage

In the third stage, experiments are carried out on comprehensive dataset of panoramic and close-range based on the networks of ResNeXt-50+FPN, ResNeXt-101+FPN and ResNeXt-50 backbone combined with Mask R-CNN. The experimental results of target segmentation are shown in Table 4.

| backbone          | mask AP | mask AP50 | mask AP75 | mask APs | mask APm | mask APl |
|-------------------|---------|-----------|-----------|----------|----------|----------|
| Mask R-CNN        | ResNeXt-50-FPN | 0.76     | 0.91      | 0.81     | 0.47     | 0.69     | 0.78     |
| Mask R-CNN        | ResNeXt-101-FPN | 0.84     | 0.95      | 0.87     | 0.60     | 0.76     | 0.84     |
| Mask R-CNN        | ResNeXt-50     | 0.66     | 0.84      | 0.72     | 0.40     | 0.59     | 0.67     |

In the third stage, two networks based on ResNeXt-50+FPN, ResNeXt-101+FPN and Mask R-CNN are used to visualize the test set images. The visualization results of the trained model on the test set are shown in Figure 6.
From the experimental results, we can see that on the comprehensive dataset of panoramic and close-range, the model based on ResNeXt-101+FPN and Mask R-CNN has the best effect. The edge of mask is smoother, and the recognition accuracy and recall rate of components are higher.

Compared with the three stages of the experiment, the test results of the three training datasets, panoramic and close-range comprehensive dataset are generally better than those of the other two datasets. From the results of visualization, we can also see that the test results of integrated dataset make the mask edge smoother and the image segmentation more accurate.

Compared with the three kinds of networks, the network based on ResNeXt-101+FPN and Mask R-CNN has the highest accuracy in three stages of dataset. In addition, FPN network is good at dealing with small targets. Compared with APs indicators, the network that increases FPN feature extraction performs better than the network that only uses ResNeXt.

5. Conclusion
In this paper, in-depth learning is introduced into the recognition of vehicle appearance components in complex scenes. The model of Mask R-CNN combined with three backbone networks is experimented on panoramic dataset, close-range dataset and panoramic close-range integrated dataset. Finally, on the comprehensive dataset of panoramic and close-range, based on ResNeXt-101+FPN and Mask R-CNN network, the training model has the best effect, the recognition accuracy is 84.4%, and the AP₅₀ is as high as 95.1%.

According to the need of practical application scenarios, this paper overcomes the difficulty of insufficient datasets, and uses data enhancement method to produce close-range dataset. Experiments show that the training set is more effective after adding close-range data.

In future research, more types of vehicle appearance components and more data sets can be added to improve the generalization ability of the model. In addition, the network structure can be further improved to improve the accuracy of the model.

References
[1] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. Mask R-CNN. In IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969.
[2] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015, pp. 91-99.
[3] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. Feature Pyramid Networks for Object Detection. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117-2125.
[4] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1492-1500.
[5] Yann Lecun, Leon Bottou, Yoshua Bengio, Patrick Haffner. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[6] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In NIPS, 2012, pp. 1097-1105.
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.
[8] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. Feature Pyramid Networks for Object Detection. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117-2125.