Vehicle Re-identification by Fusing Multiple Deep Neural Networks

Chao Cui, Nong Sang, Changxin Gao, Lei Zou
Key Laboratory of Ministry of Education for Image Processing and Intelligent Control
Huazhong University of Science and Technology
Wuhan, China
E-mail: {cuichao0819, nsang}@hust.edu.com

***Abstract*—Vehicle re-identification has become a fundamental task because of the growing explosion in the use of surveillance cameras in public security. The most widely used solution is based on license plate verification. But when facing the vehicle without a license, deck cars and other license plate information error or missing situation, vehicle searching is still a challenging problem. This paper proposed a vehicle re-identification method based on deep learning which exploit a two-branch Multi-DNN Fusion Siamese Neural Network (MFSNN) to fuses the classification outputs of color, model and pasted marks on the windshield and map them into a Euclidean space where distance can be directly used to measure the similarity of arbitrary two vehicles. In order to achieve this goal, we present a method of vehicle color identification based on Alex net, a method of vehicle model identification based on VGG net, a method of pasted marks detection and identification based on Faster R-CNN. We evaluate our MFSNN method on VehicleID dataset and in the experiment. Experiment results show that our method can achieve promising results.**

Keywords—Vehicle re-identification; MFSNN; Alex net; VGG net; Faster-RCNN

I. Introduction

Vehicle has attracted massive focuses in computer vision research field. There is a growing requirement of vehicle re-identification (Re-ID) and retrieval from large scale surveillance image and video database in public security systems. Fig. 1 gives us a description of this task. License plate is widely used as a unique ID of a vehicle, and license plate recognition has already been used in transportation management applications. However, in some cases the information of license cannot be used. Therefore, vision-based vehicle re-identification has a great practical value in real-world surveillance applications. Specifically, vehicle re-identification aims at identifying the same vehicle through different camera views.

Though the problem of vehicle re-identification has already been discussed for many years, most of the existed works rely on a various of different sensors [1]. Related works seems to be very little. There are few datasets specially designed for the vehicle re-identification task. In this paper, we used a vehicle re-identification dataset named "VehicleID" [9], which is collected from multiple real world surveillance cam-eras and includes over 200,000 images of about 26,000 vehicles. All images are attached with id numbers indicating their true identities (according to their license plate). In addition, nearly 90,000 images of 10,319 vehicles in this dataset have been labeled with the vehicle model information. Thus, it can also be used for fine-grained vehicle model recognition.

Another possible reason is that compared with the classic person re-identification problem, vehicle re-identification could be more challenging since different vehicles of one same model have similar visual appearance. It is really difficult even for humans to tell the difference between vehicles of the same model without using their license plates. Nevertheless, there are some special marks that can be used to identify a vehicle from others, such as some customized painting, favorite decorations, or even scratches etc. Therefore, Deep features should capture both the inter-class and intra-class difference efficiently. Colors and models are firstly coming up to be used for this task. For vehicles of different colors and models, it is easier for us to make a distinction between them. But for those of same colors and models, we need to use other features. In this paper, we added the feature of pasted marks on the windshield to address the vehicle re-identification problem.

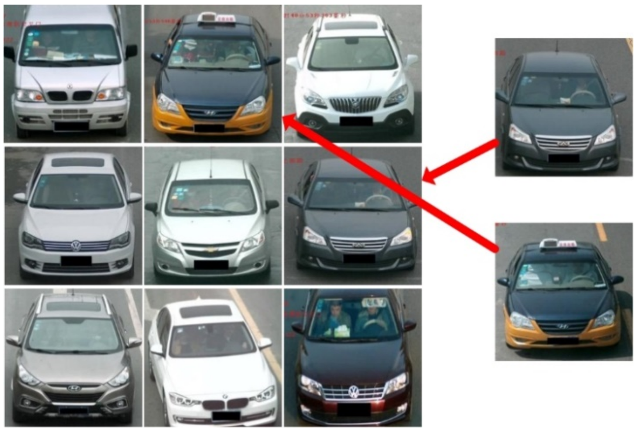


Fig. 1. Vehicle Re-identification task.

Multi-DNN Fusion Siamese Neural Network (MFSNN) is an end-to-end frame-work (Fig. 6) specially designed for vehicle re-identification. Like other Siamese Neural Network (SNN), it has two big branches for two different pictures. Each branch is spliced into three small branches, respectively, corresponding to the color; model and pasted marks feature extraction. Finally, they are merged to become a new feature.

978-1-5386-1842-4/17/$31.00 ©2017 IEEE

The experimental results show that our method of comparing this feature of each vehicle can achieve promising results and outperforms several state-of-the-art approaches.

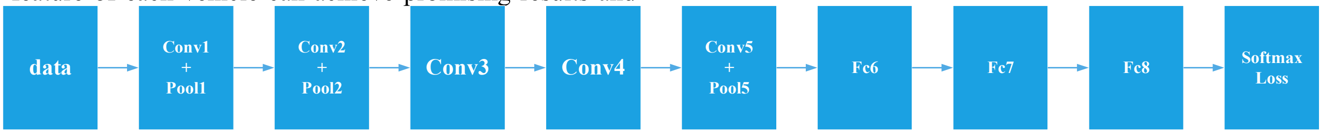


Fig. 2. Color identification network based on Alex net.

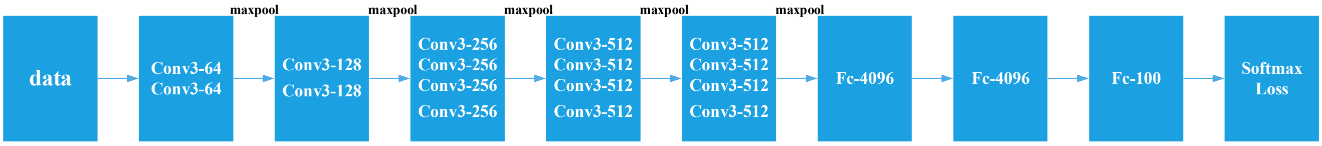


Fig. 3. Model identification network based on VGG_ILSVRC_19_layers. The convolutional layer parameters are denoted as "conv<receptive field size>-<number of channels>". The ReLU activation function is not shown for brevity.

Rest paper is organized as follows: Related works are reviewed in section II. In section III, we discuss our multi-DNN fusion based Siamese Neural Network designed for vehicle re-identification. Section IV gives a detailed description of the dataset "VehicleID" including how it is collected and organized in raw images, the total number of vehicles and extra vehicle model annotations on part of this dataset. The evaluation protocols and experimental results are presented in section V.

II. Related Work

Most previous object identification researches target at either person or human face. They can be described as a unified problem: given a probe image and multiple candidates as the gallery, we need to decide which one in gallery is the same object of the probe image. However, there is not much work on vehicle re-identification before even though vehicle is at least of equal importance as person and human face in real-world applications. The most closely related problems which targets at vehicle include vehicle model classification [12] and vehicle model verification [13]. But all those methods can only reach the vehicle model level instead of identifying whether two vehicles are exactly the same one.

There are very few articles on vehicle re-id research directly. Some research [9] draws on person re-id methods to do the vehicle task and has good results. Some method [11] mixed different features of a car and introduce spatiotemporal relation. But it is only suitable for specific dataset.

Deep convolution network is also introduced into person re-identification problem. Yi et al. [14] applied a "Siamese" deep network which has a symmetric structure with two sub-networks to learn pair-wise similarity. The two input images are first fed into two convolutional layers to extract high level features and then mixed together by a difference measurement layer together with several other fully-connected layers. The last layer in this network is a softmax function to yield the final estimate of whether the input images are of the same person

Draw on the experience of these papers, we first present some new features based on the pasted marks on the windshield, and then combine these new features with the model and color features of a vehicle. We proposed a Multi-DNN Fusion Siamese Neural Net-work (MFSNN) with two branches. Each branch has three sub networks separately to extract the features of color, model and pasted marks. We aim at making use of more features as possible to tell whether two pictures belong to the same car.

III. Multi-DNN Vehicle Re-identification

A. Vehicle color identification

Color is one of the most striking features of a vehicle. In this paper, Alex net [2] is used as the basic network structure to identify the color of a vehicle. Alex net has a very fast speed of detection, high accuracy and relatively simple structure. Thus it is able to meet our requirements.

As Fig. 2 shows, it contains 5 convolutional layers and 3 fully-connected layers. The dimension of the network's last fully-connected layer "fc8" is 7, which means 7 kinds of colors used in the dataset. We use softmax layer when training the network. The output of layer "fc8" represents the confidence probability for each color, belonging to [0, 1].

B. Vehicle model identification

Compared with color identification, model identification is more difficult. Thus we used a network with higher accuracy and more complex structure: VGG_ILSVRC_19_layers [3]. Its complex structure is showed in Fig. 3. The dimension of the network's last fully-connected layer "fc-100" is 100, which means 100 kinds of models used in the dataset for training. The output of layer "fc-100" represents the confidence probability for each model, belonging to [0, 1].

C. Pasted marks detection and identification

There is a small but quite important difference between identifying a specific vehicle and person. In theory, any two persons could not be exactly the same regarding their visual appearance but two vehicles running on road could be if they belong to the same color and vehicle model. But in real-world scenes, it is still possible to distinguish two vehicles of the same model if some special markers exist. To deal with this case, we need to use some unique feature of a car.

In this paper, we select the pasted marks on the front windshield as newer non-license information of a car. Through large amount of observation and query induction, we concluded the following characteristics (as Fig. 4 shows) of pasted marks: the number of signs (mainly within 4); the internal different distribution and the overall distribution in the upper right corner of the car; limited color Types (mainly green, yellow and white); similar area size. So the number, content, arrangement and distribution of marks become our focusing point. The number is easy to obtain. The content of the image (due to the quality limit) can only be represented by the color information. Arrangement and distribution information in this paper is obtained by the relative position of the marks and the windshield. As shown in Fig. 4, according to the characteristics of the mark with the number, color, and the front glass relative position as the chapter selected paste mark characteristics. The method is as follows: the number of detected markers is used as the quantity information; the color of the detected adhesive is used as the color information; the mean tangent between the connections which is between the center of the detected adhesive and that of the front glass, and the horizontal line is used as the relative position information.

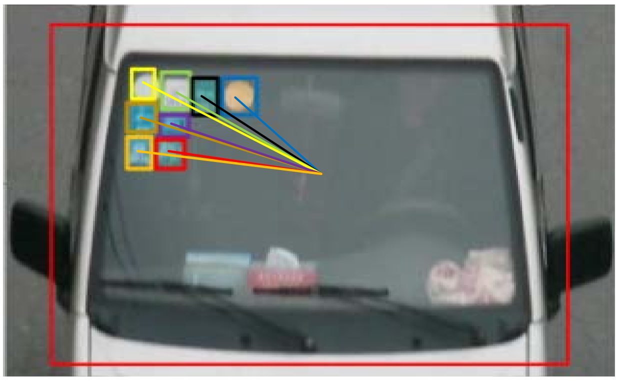


Fig. 4. Mark centers and the windshield center connection diagram

For the color information, the RGB [4] value and the HSL [5] value for each pixel has a single color, and the individual values of the mark are converted to gray value, the conversion formula is:

$$Gray = R\times 0.299 + G\times 0587 + B\times 0.114 \quad (1)$$

Conversion relationship between H and RGB is:

$$M = max(R, G, B) \quad (2)$$

$$m = min(R, G, B) \quad (3)$$

$$C = M - m \quad (4)$$

$$H' = \begin{cases} 0, (C = 0) \\ \dfrac{G-B}{C}\%6, (M = R) \\ \dfrac{B-R}{C} + 2, (M = G) \\ \dfrac{R-G}{C} + 4, (M = B) \end{cases} \quad (5)$$

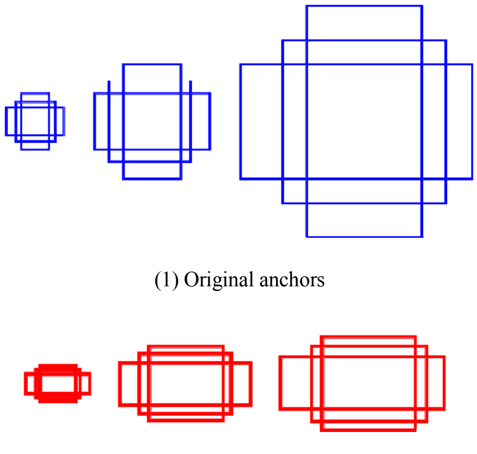
$$H = 60°\times H' \quad (6)$$

Thus we can get the gray mean and H mean. The gray mean, H mean and tangent mean features are achieved by improving the Faster R-CNN [6] Network implementation. Faster R-CNN can be seen as RPN + Fast R-CNN, with regional generation network (RPN) instead of Fast R-CNN in the candidate area selection. One part of the overall network achieves a rough detection area that is candidate area generation network (RPN), the other part completes the classification of the target, the two share part of the network structure.

After the input of the picture, the whole recognition is done by three parts. Firstly, the image-based candidate region is generated. The generation of the region is cut by the picture and does not depend on the category to which the object belongs in the image. Secondly, the feature vector is extracted from the candidate region. Finally, the feature vector is sent to the classifier for each category.

The distribution of the paste marks, the aspect ratio and the area size are fixed. Thus there is a certain constraint relationship, which will be improved on the basis of Faster-R-CNN to increase the detection accuracy of the network for the mark features. We adjust the anchors of the region-generated network, based on the size of the marks.

After attempts of different anchor sizes, we found that the close size of the mark and the anchor, the similar ratio of the length and width make the mark detection more accurate. The original anchor sizes of the network are three kinds: 128 × 128, 256 × 256, 512 × 512, with 3 aspect ratio (1: 1, 1: 2, 2: 1) combinations, a total of 3 * 3 = 9 kinds. We analyzed the size and aspect ratio of the mark statistically, and then achieve the result that the average area size is 34 × 25 and the aspect ratio is mainly 3: 2. The aspect ratio of the paste mark and the anchor is not similar, neither is the size. So we modify the anchor, taking into account the size of the marks and the windshield. The original size and the modified size of the anchor is showed in Fig. 5. The modified anchor sizes of the network are three kinds: 60 × 60, 120 × 120 and 150× 150, with 3 aspect ratio (3: 1, 3: 2, 3: 3) combinations. After the modification, we are able to precisely detect the features of the pasted marks.



(1) Original anchors

(2) Modified anchors

Fig. 5. Size of original anchors and modified anchors

The output of the network includes the number of marks which belongs to [0, 10], grey mean which belongs to [0, 255], H mean which belongs to [0, 360], tangent mean which belongs to [0.577, 1.732] (counting by us). Each of the features is normalized to [0, 1].

D. Multi-DNN Fusion Siamese Neural Network

To do the re-id task, we designed a Siamese Neural Network (SNN) [7]. The main ideal of the SNN is to learn a function that maps input patterns into a latent space, in which the similarity metric will be large for same objects, and small for pairs from different ones. Therefore, it is best suited for re-id task. In order to judge whether two cars from different pictures are the same one, we use two pictures as an input pair for the two-branch pair. Each branch extracts the features of a car. It fuses the classification outputs of color, model and the features of the pasted marks on the windshield. Thus each branch is divided into three small branches then concatenated into a fully-connected layer “fc_final”. The dimension of layer “fc_final” is 111.The structure of the network and the output dimensions of each layer are shown in Fig. 6.

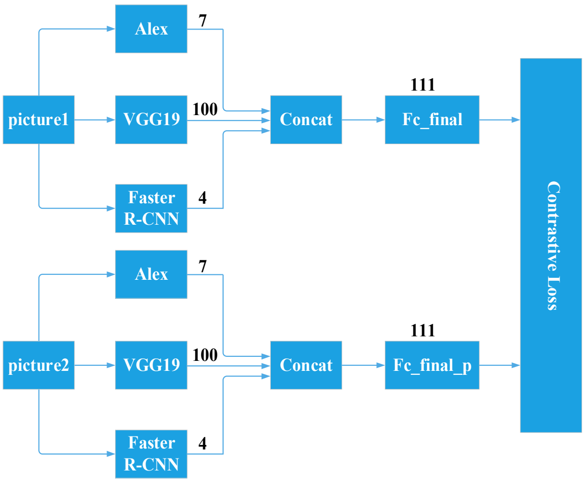


Fig. 6. Structure of Multi-DNN Fusion Siamese Neural Network (MFSNN)

During training, the pairwise images are fed into the two CNNs separately. After the forward propagation, the outputs of CNNs are combined into the contrastive loss layer to compute the loss of the model. Then through the back propagation with the contrastive loss, the shared weights of the two CNNs are optimized simultaneously. Given a pair of car images, we can map the data into the latent metric space to measure the compatibility between two pictures.

Specifically, let W be the weights of SNN, given a pair of car images x_1 and x_2 we can map the data into the latent metric space as $S_w(x_1)$ and $S_w(x_2)$. Then the energy function, $E_w(x_1,x_2)$, which measures the compatibility between x_1 and x_2, is defined as

$$E_w(x_1,x_2)=\|S_w(x_1)-S_w(x_2)\| \tag{7}$$

With the energy function, the contrastive loss can be formulated as

$$L(W,(x_1,x_2,y))=(1-y)\cdot\max(m-E_w(x_1,x_2),0)+y\cdot E_w(x_1,x_2) \tag{8}$$

Where (x_1 , x_2 , y) is a pair of samples with the label, m is a positive margin. During the test, we using a the learned SNN to extract a 111-D output from a concat layer. The Euclidean distance is adopted to estimate the similarity of the pair images.

E. Train the net

All the networks in our experiments are trained with the widely-used deep learning framework “Caffe” [8]. Training data consist of positive and negative pair images with multiple labels, including color, model, coordinates of the pasted marks and vehicle Id.

If we directly train the MFSNN, it will be too difficult to make the network converge because of its complex structure and parameters. Thus we separately train the Alex net for color, VGG_ILSVRC_19_layers for model and Faster R-CNN for pasted marks then load their parameters in MFSNN. In the Implementation, we adopt the contrastive loss layer with the default margin value 1. We use a momentum of μ = 0.9 and weight decay λ = 0.0001. Batch-size is set to 15 which means we need to feed 15 pairs of pictures in each training iteration. We start with a base learning rate of η = 0.01 and then drops by repeatedly multiply 0.7 after every 4000 batches iterations. When training the sub-branch network, we use the same parameters.

IV. VehicleID Dataset

As mentioned in section 1, to do the vehicle re-id task, we use a suitable large-scale vehicle dataset named “VehicleID”. It has been carefully collected organized by Hongye Liu et al. [11] The “VehicleID” dataset contains data captured during daytime by multiple real-world surveillance cameras distributed in a small city in China. Similar to existed person re-identification datasets, each vehicle ever appeared includes more than one image. Thus, this dataset could be well suitable for vehicle search related tasks.

In addition, they carefully labeled 10319 vehicles (90196 images in total) of their vehicle model information. But different from the “CompCars” dataset [10], the dataset does not target at fine-grained vehicle model classification task since the model distribution is usually quite imbalanced in real-world scenarios. In “VehicleID” dataset, only 250 most commonly appeared vehicle models are included (like “MINI-cooper”, “Audi A6L” and “BWM 1 Series”).

The “VehicleID” dataset is captured by multiple non-overlapping surveillance cameras and there are 221763 images of 26267 vehicles in total (8.44 images/vehicle in average). Besides, the vehicle in each image is either captured from the front or the back (viewpoints information is not labeled). Fig. 7 demonstrates some examples.

In this paper we eliminate the number of sample models less than 300, and then pick 43075 pictures with 100 models and 7 colors. Each car has its unique vehicle ID. All the license plates are added mosaic on to eliminate the effects of the license number. Then we picked 1000 pictures of 200 cars to label the coordinates of the front windshield and pasted marks. Training sets and test sets are divided according to 8:2.

As mentioned in section III, we separately trained the Alex net for color, VGG_ILSVRC_19_layers for model and Faster R-CNN for pasted marks. We designed two experiments, first of which was to evaluate our modification on the anchor size of Faster R-CNN network. The second experiment compared different methods on the vehicle retrieval task. The detailed description and final results are in the following two subsections.

Fig. **7.** Samples of "VehicleID" dataset

V. Experiments

A. *Pasted Marks Detection*

In section III we give a modified Faster R-CNN network with changed size of anchor. In this experiment we comprehensively evaluate our method of changed size of anchor on the "VehicleID" dataset to see if it is valid. Thus we separately trained the original Faster R-CNN network and the network after changing the size of anchor, using the train set of 800 pictures and tested it with the test set which contained 200 pictures. We set the IoU (Intersection-over-Union) for NMS (Non Maximum Suppression) at 0.7, just as the author in [6] did. The detection accuracy is shown in Table 1.

From the table it is obviously drawn that the change of anchor size will make significant effects of pasted marks detection. A sample tested picture and its detection result is given in Fig. 8.

Table 1. Accuracy of detection task

	Train number	Test number	Accuracy before resize	Accuracy after resize
Marks	800	200	0.171	0.705

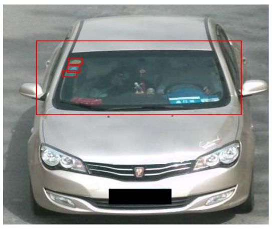


Fig. 8. Sample detection result

B. *Vehicle Re-identification*

Vehicle re-identification is our mainly task in this paper. We evaluate the performance of our proposed MFSNN model following the widely used protocol in object retrieval, mean average precision (MAP). We designed this experiment to measure how much improvement each module in our framework brings and to compare our method with others. We compared our three kinds of Siamese Neural Network in this part: two branches of Alex net and VGG_ILSVRC_19_layers using color and model features, two branches of Faster R-CNN net only using pasted mark features and MFSNN combining them all. All of these three methods extract the normalized features using the trained deep convolutional network and after then the difference arbitrary two vehicle images is measured directly by their L2 distance. Three other methods are introduced to perform the comparison experiment. The experiment results of these three methods, "VGG+Triplet Loss", "VGG+Coupled Clusters Loss" and "Mixed Difference VGG+Coupled Clusters Loss" are referred from Liu's paper [11].

Considering the total number of testing data is too large compared with ordinary testing data for person re-identification (316 pedestrians in VIPeR dataset, 50 in iLIDS dataset, we extract three subsets (i.e. small, medium, large) ordered by their size from original testing data for the re-identification task. The quantity distribution of "VehicleID" is demonstrated on Table 2.

Table 2. Test Data Split

	Small	Medium	Large
Number of vehicles	800	1600	2400
Number of images	7210	14323	21542

Table 3 illustrates the final results. In all three testing datasets, the mean average precision keeps growing significantly after using features of color, model and pasted marks, compared with the two- branch Siamese network, each branch has only one or two small branches extracting features of color and model, or pasted marks. Thus it strongly proves the significant effects of fusing multiple classification outputs. Compared our MFSNN network with the other three methods, despite its MAP is lower than the "Mixed Difference VGG+Coupled Clusters Loss", it still gets better results in the three testing datasets than the other two networks. Thus the results are acceptable and feasible.

Table 3. MAP of Vehicle Re-identification Task

MAP	Small	Medium	Large
VGG+Triplet Loss	0.444	0.391	0.373
VGG+CCL	0.492	0.448	0.386
Mixed Diff+CCL	0.546	0.481	0.455
Color and Model	0.401	0.389	0.351
Pasted marks	0.473	0.429	0.384
MFSNN	0.527	0.474	0.427

VI. Conclusions

In this paper, we proposed a Multi-DNN Fusion Siamese Neural Network (MFSNN) to solve an important but not well explored problem: vehicle re-identification. We exploit a two-branch deep convolutional network to map the fused classification outputs of color, model and pasted marks on windshield into a Euclidean space to tell whether two pictures belong to the same car. Beside the common features (color and model), we provide some new features of the pasted marks to solve the problem. Through fusing the classification outputs of common features and our newly exploited features, our proposed method achieves promising results. Compared with existed methods, the specifically designed network structure achieves a high predict accuracy. Experimental results demonstrate that after changing the size of anchor we can get higher detection accuracy on pasted marks of the vehicle windshield. And it also has been drawn that our MFSNN outperforms several state-of-the-art approaches on the vehicle re-identification task.

References

[1] C. C. Sun, G. S. Arr, R. P. Ramachandran, and S. G. Ritchie. Vehicle reidentification using multidetector fusion. Intelligent Transportation Systems, IEEE Transactions on, 5(3):155–164, 2004.

[2] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C].in: IEEE International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531, 2014.

[4] R.C. Gonzalez , and R. E. Woods, Digital Image Processing 2nd Edition. Prentice Hall, 2002.

[5] S. Westland, and A. Ripamonti, Computational Colour Science. John Wiley, 2004.

[6] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, PP (99):1-1.

[7] Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., S¨ackinger, E., Shah, R.: Signature verification using a siamese time delay neural network. Int. J. Pattern Recogn. Artif. Intell. 7(04), 669–688 (1993).

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

[9] Liu, Hongye, et al. "Deep Relative Distance Learning: Tell the Difference between Similar Vehicles." IEEE Conference on Computer Vision and Pattern Recognition IEEE, 2016:2167-2175.

[10] L. Yang, P. Luo, C. C. Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3973–3981, 2015.

[11] Liu, Xinchen, et al. "A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance." European Conference on Computer Vision Springer International Publishing, 2016:869-884.

[12] Z. Zhang, T. Tan, K. Huang, and Y. Wang. Three dimensional deformable-model-based localization and recognition of road vehicles. Image Processing, IEEE Transactions on, 21(1):1–13, 2012.

[13] L. Yang, P. Luo, C. C. Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3973–3981, 2015.

[14] Yi D, Lei Z, Li S Z. Deep Metric Learning for Practical Person Re-Identification[J]. Computer Science, 2014:34-39.