# MULTI-ATTRIBUTE DRIVEN VEHICLE RE-IDENTIFICATION WITH SPATIAL-TEMPORAL RE-RANKING

*Na Jiang, Yue Xu, Zhong Zhou *, Wei Wu**

State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China

## ABSTRACT

Vehicle re-identification (re-id) is a promising topic, which focuses on retrieving the same vehicles across different cameras. It is challenging due to the variations of illumination and camera viewpoints. To solve these problems, we present a multi-attribute driven vehicle re-id approach to learn discriminative representations. The proposed approach consists of a multi-branch architecture and a re-ranking strategy. The multi-branch architecture extracts color, model, and appearance features, which explicitly leverages the vehicle attribute cues to enhance the generalization ability, especially for the different vehicles with similar appearance and the same vehicles with different orientations. The re-ranking strategy introduces the spatial-temporal relationship among vehicles from multiple cameras to construct the similar appearance sets and utilizes Jaccard distance between these similar appearance sets to re-rank. Extensive experimental results demonstrate that our proposed approach significantly outperforms state-of-the-art re-id methods on the popular VeRi-776 dataset and VehicleID dataset.

***Index Terms***— Vehicle Re-Identification, multi-attribute driven architecture, spatial-temporal re-ranking.

## 1. INTRODUCTION

Vehicle re-identification (re-id) refers to retrieving the same vehicles from large-scale surveillance videos. In current applications of traffic management, license plate recognition plays an important role to provide vehicle identify for vehicle re-id. However, it is difficult to capture license plates in many surveillance cameras due to viewpoint, occlusion and illumination. What's more, the criminals often use fake license plates to escape search. Vehicle re-id bases on appearance features, therefore, becomes the effective way to solve such problems and quickly becomes a research focus in field of computer vision. It is challenging due to the orientation variations, illumination changes, and amounts of similar vehicle models. As shown in Fig. 1, Fig.1 (a) represents the same vehicle model with different color, Fig.1 (b) shows the two different vehicles with similar color and model, Fig.1(c) and Fig.1 (d) are the same vehicle identify with different orientations. In the face of these images, the performance of traditional methods have dramatic decline, and even the observation ability of humans is also prone to generate the wrong classifications.

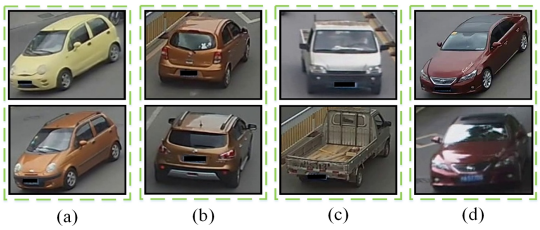



**Fig. 1**. Some hard examples of vehicle re-identification

Inspired by the person re-id [1, 2, 3], some success deep learning frameworks [4, 5, 6, 7] that have make great breakthroughs in computer vision are introduced into vehicle re-id to extract appearance features[8, 9]. However, these existing methods still often fail in the cases displayed in Fig.1. The main reason is that they only focus on extracting appearance features [10, 11, 12], while ignore to explore the attribute cues and spatial-temporal relationships of vehicles. To make up for this deficiency, we proposed a novel multi-attribute architecture that exploits attributes to improve the feature representations of raw vehicle images.

The architecture consists of a backbone network and two branch networks. The backbone is responsible for extracting appearance features. The two branches extract color and vehicle model features respectively. Two attribute branches drive the backbone to extract more discriminative representations, which can alleviate the variations of illumination and orientation. Furthermore, we also design a spatial-temporal re-ranking strategy to optimize the vehicle re-id algorithm. The strategy makes full use of the spatial-temporal relationship between every image pairs to generate similar appearance sets for re-ranking by Jaccard distance. To evaluate the proposed method, we conduct multi-group comparative analysis experiments on VeRi-776 dataset and VehicleID dataset. Experimental results demonstrate that our method outperforms most

*Corresponding author {zz, wuwei}@buaa.edu.cn

This work is supported by the Natural Science Foundation of China under Grant No. 61472020, 61572061, 61502020, and the China Postdoctoral Science Foundation under Grant No. 2013M540039.

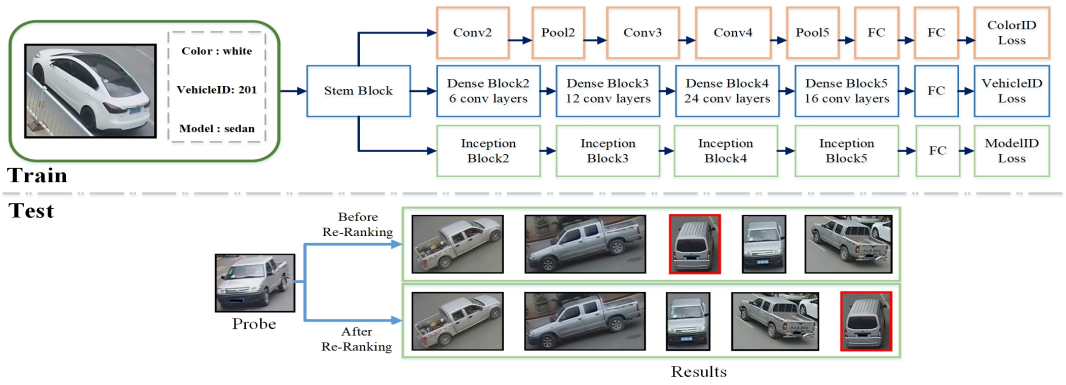



**Fig. 2**. Outline of our proposed method

state-of-the-art methods. On VeRi-776 dataset, our proposed method achieves 2.84% improvements in mAP and 5.71% in Rank-1. On VehicleID dataset, we also get various degrees of improvements for three different scale subsets.

## 2. OUR APPROACH

In this section, we demonstrate the outline of our proposed method in Fig.2. Meanwhile, we also introduce the multi-attribute architecture designed for learning feature representations and the spatial-temporal re-ranking strategy.

### 2.1. Multi-Attribute Architecture

Existing methods of person re-identification and vehicle re-identification usually extract appearance features for similarity metric. Although appearance features extracted from excellent deep learning frameworks can describe the majority of the appearance information, erroneous classifications are still inevitable. Inspired by multi-branch networks from person re-id, we find that extracting various features can improve the discrimination and robustness of feature representations. We immediately think of the vehicle model attribute to drive the backbone network. Different from non-rigid persons, vehicles are rigid objects whose shapes and models are very stable. The model attribute can be employed to learn the corresponding relationships of the same vehicle model with different orientations. Unfortunately, model attribute features are easily to lead to the incorrectly associations displayed in Fig.1 (a), due to ignoring the color. Therefore, the color attribute should also be consider for vehicle re-id.

As shown in Fig.2,we propose a multi-attribute architecture with different convolutional blocks to learn different labels. It consists of a stem block, a backbone network and two branch networks. The stem block contains two normal CNN layers and one Pooling layer. They share the parameter weights, which is beneficial to back propagation and reduce computing resources. Due to different labels, the backbone network has different convolutional blocks with two attribute branches. Analysis the existing deep learning structures, we find that various deep learning structures have different sensitivities to every attributes. The simple CaffeNet is sensitive to the colors and the GoogleNet[5] is more concerned about the vehicle model attributes. The latest DenseNet [7] can achieve discriminant appearance features by reusing low-level features. To this end, we exploit dense blocks to design the backbone network, modify inception blocks to build the model branch network, and choose CaffeNet as the color branch network. The details of multi-attribute architecture are described in Table 1.

**Table 1**. Sketch map of our network structure

| Branch | Module Name | Output Size |
|---|---|---|
| Appearance | Stem Block | 64x57x57 |
| | Dense block2 | 128x29x29 |
| | Dense block3 | 256x15x15 |
| | Dense block4 | 512x8x8 |
| | Dense block5 | 1024x1x1 |
| | FC | 512x1x1 |
| Color | Conv2 | 256x57x57 |
| | pool2 | 256x29x29 |
| | Conv3 | 384x29x29 |
| | Conv4 | 384x15x15 |
| | Conv5 | 256x8x8 |
| | Pool5 | 256x1x1 |
| | FC6 | 512x1x1 |
| | FC7 | 512x1x1 |
| Model | Inception Block2 | 256x57x57 |
| | Inception Block3 | 256x29x29 |
| | Inception Block4 | 512x14x14 |
| | Inception Block5 | 1024x7x7 |
| | Pool | 1024x1x1 |
| | FC | 512x1x1 |

In the training phase, we adopt two-stage training strategy. In the first stage, three softmax loss functions are used to train the backbone and two branch networks. In the second stage, we exploit softmax loss function, improved triplet loss function, and hard example mining strategy [13] to jointly train the backbone network. Compared with the original

triplet loss function, our improved triplet loss function introduces the intra-class constraint.The original triplet loss function $L_{id}(I_i^a, I_i^p, I_i^n)$ is defined as follow:

$$L_{id}(I_i^a, I_i^p, I_i^n) = \sum_{i=1}^{N}[d(f(I_i^a), f(I_i^p)) - d(f(I_i^a), f(I_i^n)) + \alpha]_+ \quad (1)$$

where $I_i^a$ represents the anchor image in a triplet input, $I_i^p$ denotes the positive sample of anchor image, $I_i^n$ expresses the negative sample of the anchor image.$[x]_+ = max(x, 0)$. $d(x, y)$ represents the L2-norm distance between $x$ and $y$. $\alpha$ is a margin between positive and negative samples, and $N$ is the number of triples. On this basis, the introduced intra-class constraint is defined as follows:

$$L_{in}(I_i^a, I_i^p) = \sum_{i=1}^{N}[d(f(I_i^a), f(I_i^p)) + \beta]_+ \quad (2)$$

where $\beta$ is a threshold of intra-class similarity constraint. The improvement makes the features from different cars farther away from each other, meanwhile features from the same car closer.

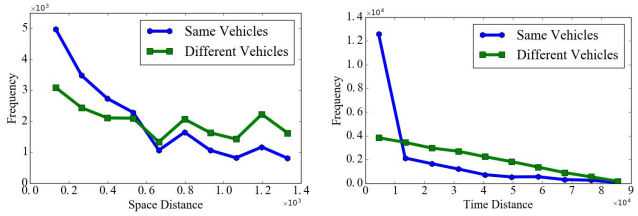
In the testing phrase, the achieved color and model attribute features are set as the filter conditions. Only when an image has similar attributes with the probe image, we calculate their appearance feature distances for vehicle re-id. This retrieval mode implicitly improves the rank-$k$ accuracy of our proposed algorithm by reducing the number of the gallery set.

### 2.2. Spatial-Temporal Re-Ranking

In the real monitoring environments, each independent vehicle image not only provides appearance information, but also provides spatial-temporal information. Attribute driven appearance features only describe the appearance information of each independent vehicle image. To exploit the spatial-temporal information, we explore the spatial-temporal relationships between every image pair. As shown in Fig.3, we draw the statistics of spatial-temporal information from VeRi-776 dataset. The blue lines represent the statistical results of the same vehicles, and the green lines denote the results of random different vehicles. It is obvious that the number of the same vehicles with small space or time distance is more than the same vehicles with large space or time distance.

According to the observation, we design a spatial-temporal re-ranking strategy to further optimize the vehicle re-id. Firstly, we achieve the initial retrieval results by Euclidean distance. Then, we define $S(I, k)$ as the top-$k$ similar appearance set of image $I$, which depends on the associated score. The associated score considers the spatial-temporal relationship and the appearance similarity simultaneously. It can be calculated by Eq.3.

$$C(I_i, I_j) = \frac{\|T_i - T_j\|}{T_{max}} \times \frac{\delta(D_i - D_j)}{D_{max}} \times d(f(I_i), f(I_j)) \quad (3)$$




(a) (b)

**Fig. 3**. Statistics of spatial-temporal information

where $T_i$ and $T_j$ are the timestamps of $I_i$ and $I_j$. $T_{max}$ is the max time difference between every probe image and gallery images. $D_{max}$ denotes the max spatial distance among all cameras provided images. $\delta(D_i - D_j)$ is the smallest distance between $I_i$ and $I_j$. The image pairs with small associated score have high spatial-temporal relationship and appearance similarity.

Following Eq.3, we can achieve the similar appearance sets of the probe image and retrieval results. Taking them as inputs, the distance between the probe image and the retrieval results can be re-calculated by Jaccard distance. It is defined as follow:

$$d_J(p, r_i) = 1 - \frac{S(p, k) \bigcap S(r_i, k)}{S(p, k) \bigcup S(r_i, k)} \quad (4)$$

where $S(p, k)$ denotes the similar appareance set of the probe image $p$, $S(r_i, k)$ represents the similar appareance set of the results $r_i$. In this paper, the $k$ is empirically set to 6.

## 3. EXPERIMENTS

To verify the proposed algorithm, we conduct comparative experiments on VeRi-776 dataset[8] and VehicleID dataset[11]. The implement details of experiments are described in Sec. 3.1 and the results are demonstrated in Sec. 3.2.

### 3.1. Implement Details

The selected VeRi-776 dataset provides 37781 images for training and 13257 images for testing. VehicleID dataset contains data captured during daytime by multiple real-world surveillance cameras distributed in a small city in China. There are 26267 vehicles in the entire dataset. Due to the ratio imbalance between different classes, we eliminate the vehicles with fewer than six images. In the comparative experiments, we perform our algorithm on Caffe platform and define a data layer using Python interface to support multi-attribute inputs. For data augmentation, we resize all images to 256*256 and crop them into 224*224 with horizontal flip. During the joint training process, initial learn rate is set to 0.01 and maximum number of iterations is set to 30 epoches. At the evaluation stage, we adopt mean average precision

(mAP) and Rank-1 to compare our proposed algorithm with the state-of-the-art approaches.

### 3.2. Analysis of Contribution Effectiveness

In this section, we design the following experiments on VeRi-776 dataset to analyze the effectiveness of contributions. The experimental results from different optimizations are shown in Table 2.

**Table 2**. Analysis results of our method with different settings

| Setting | mAP(%) |
|---|---|
| Appearance | 54.94 |
| Appearance+Color | 56.92 |
| Appearance+Color+Model | 58.05 |
| Appearance+Color+Model+Re-Ranking | **61.11** |

In Table 2, appearance represents that results are from the backbone structure. Based on it, we add the color branch, model branch, and re-ranking strategy one by one. As is shown in Table 2, the color features and the model features improve the mAP by 1.98% and 3.11% compared with the appearance results, respectively. Meanwhile, the spatial-temporal re-ranking strategy further improves the algorithm performance. The effectiveness of contributions can also be seen in Cumulative Match Characteristic (CMC) curves (see Fig. 4).

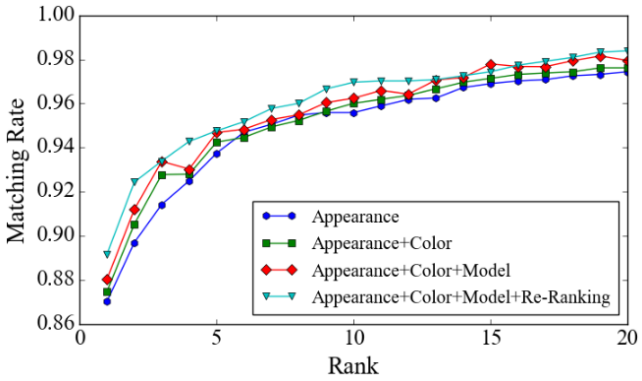



**Fig. 4**. CMC on VeRi-776 Dataset

As shown in Fig.4, our proposed complete architecture with re-ranking strategy achieves the best performance, which proves the contributions of attribute branches and spatial-temporal relationship again.

### 3.3. Performance Comparison on Popular Datasets

The proposed method is compared with recent state-of-the-art algorithms on two popular datasets. The experiments are repeated for 10 times and the average results are described in Table 3 and Table 4.

**Table 3**. Comparison with state-of-the-art approaches on VeRi-776 dataset

| Method | mAP(%) | Rank-1(%) | Rank-5(%) |
|---|---|---|---|
| FACT[8] | 18.49 | 50.95 | 73.48 |
| FACT+Plate-SNN-STR[8] | 27.77 | 61.44 | 78.78 |
| Siamese-CNN[9] | 54.21 | 79.32 | 88.92 |
| Siamese-CNN-Path-LSTM[9] | 58.27 | 83.49 | 90.04 |
| **Our Method** | **61.11** | **89.27** | **94.76** |

In Table 3, the FACT and FACT+Plate-SNN-STR combine appearance features learned by GoogLeNet, SIFT texture features and color features extracted by Color Name(CN) model. The Siamese-CNN not only extracts the appearance features from CNN framework with Siamese loss function, but also exploits the license plate information. On the basis, the Siamese-CNN+Path-LSTM introduces LSTM units to utilize the spatial-temporal paths. Even so, compared with them, our proposed method still achieves excellent performance on both Rank-1, Rank-5 and mAP.

**Table 4**. Comparative results of mAP on VehicleID dataset

| Method | Small | Medium | Large |
|---|---|---|---|
| Mix Diff+CCL[11] | 0.546 | 0.481 | 0.455 |
| HDC+Contractive[14] | 0.655 | 0.631 | 0.575 |
| DJDL[15] | 0.786 | 0.747 | 0.720 |
| **Our Method** | **0.820** | **0.759** | **0.728** |

As shown in Table 4, we also perform contrast experiment on VehicleID dataset with different scales testing set. Although [11, 15] exploit multiple loss functions to jointly train network and [14] uses hard-aware cascade to enhance network, our method still outperforms them on mAP. It is worth noting that VehicleID dataset does not provide the path information. Our performance improvement displayed in the Table 4 only owes to the combination of attribute and appearance features, which strongly proves the validity of the multi-attribute architecture.

## 4. CONCLUSION

In this paper, we proposed a multi-attribute deep learning architecture and a spatial-temporal re-ranking strategy for vehicle re-id. The architecture can extract appearance features, color features, and model features to improve the discriminative representations of raw vehicle images, which effectively incorporates the vehicle attribute information to distinguish similar vehicles. The re-ranking strategy introduces time and location to calculate the spatial-temporal relationships between vehicle pairs, which can assist appearance feature distances to re-rank the retrieval results. Extensive experiments and analysis on VeRi-776 dataset and VehicleID dataset demonstrate that our proposed approach is superior to most state-of-the-art methods. In future work, we will pay attention to vehicle re-id based on video sequences.

## 5. REFERENCES


[1] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[2] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[3] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[4] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[7] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," in *Computer Vision and Pattern Recognition*, 2017, vol. 1, p. 3.

[8] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.

[9] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," *arXiv preprint arXiv:1708.03918*, 2017.

[10] Dominik Zapletal and Adam Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1568–1574.

[11] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[12] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 3652–3661.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[14] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang, "Hard-aware deeply cascaded embedding," *CoRR, abs/1611.05720*, vol. 1, 2016.

[15] Yuqi Li, Yanghao Li, Hongfei Yan, and Jiaying Liu, "Deep joint discriminative learning for vehicle re-identification and retrieval," in *International Conference on Image Proceeding*. IEEE, 2017.