Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles

Hongye Liu[1,3], Yonghong Tian[1,3*], Yaowei Wang[2*], Lu Pang[1,3], Tiejun Huang[1,3]
[1]National Engineering Laboratory for Video Technology, Peking University, Beijing
[2]School of Information and Electronics, Beijing Institute of Technology, Beijing
[3]Cooperative Medianet Innovation Center, China
{hongyeliu, yhtian, panglu, tjhuang}@pku.edu.cn, yaoweiwang@bit.edu.cn

Abstract

The growing explosion in the use of surveillance cameras in public security highlights the importance of vehicle search from a large-scale image or video database. However, compared with person re-identification or face recognition, vehicle search problem has long been neglected by researchers in vision community. This paper focuses on an interesting but challenging problem, vehicle re-identification (a.k.a precise vehicle search). We propose a Deep Relative Distance Learning (DRDL) method which exploits a two-branch deep convolutional network to project raw vehicle images into an Euclidean space where distance can be directly used to measure the similarity of arbitrary two vehicles. To further facilitate the future research on this problem, we also present a carefully-organized large-scale image database "VehicleID", which includes multiple images of the same vehicle captured by different real-world cameras in a city. We evaluate our DRDL method on our VehicleID dataset and another recently-released vehicle model classification dataset "CompCars" in three sets of experiments: vehicle re-identification, vehicle model verification and vehicle retrieval. Experimental results show that our method can achieve promising results and outperforms several state-of-the-art approaches.

1. Introduction

Nowadays, there is an explosive growing requirement of vehicle search and re-identification (Re-ID) from large-scale surveillance image and video database in public security systems. License plate naturally is an unique ID of a vehicle, and license plate recognition has already been used widely in transportation management applications. Unfortunately, we can not identify a vehicle simply by its plate in some cases. First, most surveillance cameras are not installed for license plate capturing, thus, plate recognition performance drops dramatically on images/video data captured by these cameras. Furthermore, license plates are often occluded, removed, or even faked in a large number of previous security events. Therefore, vision-based vehicle re-identification has a great practical value in real-world surveillance applications. Specifically, vehicle re-identification is the problem of identifying the same vehicle across different surveillance camera views. Fig. 1 gives a straightforward description of it.

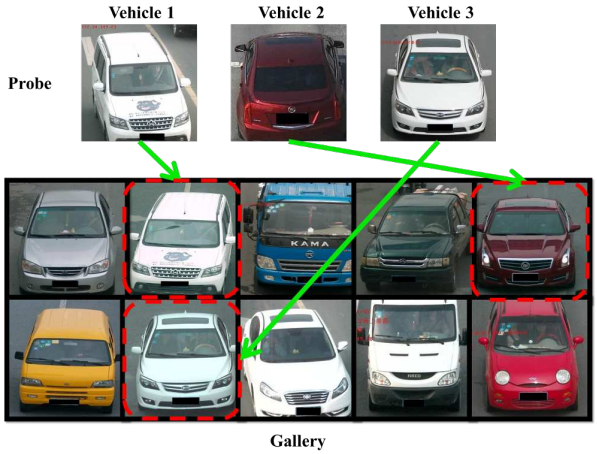

Figure 1. Given multiple candidates as the gallery set, the vehicle re-identification task is to find the matched one for each probe image. Notice that the illumination and viewpoint in different cameras can be varied a lot and different vehicles could be quite similar if they are of the same model.

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 [18, 10, 12]. To our knowledge, there is no previous attempt on the vehicle re-identification task purely by vehicle's visual appearance yet and the primary reason could be the lack of high-quality and large-scale vehicle Re-ID datasets. Existed vehicle datasets [9, 21] are usually designed for vehicle at-

*Corresponding author:Yonghong Tian and Yaowei Wang (email: yhtian@pku.edu.cn, yaoweiwang@bit.edu.cn).

tributes recognition(e.g. color, type, make, and model). In this paper, we present a new vehicle re-identification dataset named "VehicleID", which is collected from multiple real-world surveillance cameras and includes over 200,000 images of about 26,000 vehicles. All images are attached with id numbers indicating their true identities(according to the vehicle's 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 potential reason may be that compared with the classic person re-identification problem, vehicle re-identification could be more challenging as too many (usually thousands) 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. (as illustrating in Fig. 2). Therefore, vehicle re-identification algorithm should be able to capture both the inter-class and intra-class difference efficiently. Deep feature has been proved more effective and robust for recognition task. Inspired by one of the state-of-the-art method in person re-identification [4], we propose a Deep Relative Distance Learning (DRDL) model to address the vehicle re-identification problem.

DRDL is an end-to-end framework (Fig. 3) specifically designed for vehicle re-identification. It aims to learn a deep convolutional network that can project raw vehicle images into an Euclidean space where the $L2$ distance can thus be used directly to measure the similarity of arbitrary two vehicles. The basic idea of DRDL is to minimize the distances of the same vehicle images and maximize those of other vehicles. Therefore, a *coupled cluster loss* function and a *mixed difference network structure* are introduced in DRDL framework. As shown in Fig. 3, the input of DRDL are two image sets: one positive set (images of the same vehicle identity) and one negative set (images of other vehicles). The coupled cluster loss is to pull the positive images closer and push those negative ones far away. While the mixed difference network structure will benefit the mapping model with more explicit model information. Namely, deep feature and the distance metric are learned simultaneously in an unified DRDL framework. The experimental results show that our method can achieve promising results and outperforms several state-of-the-art approaches.

Rest of the paper is organized as follows: Related works are reviewed in section 2. In section 3, we discuss our coupled clusters loss function and a unified deep network structure specifically designed for vehicle re-identification. Section 4 gives an detailed description of our dataset "VehicleID" including how we collect and organize the 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 5.

Figure 2. Special marks which can be used for identification task.

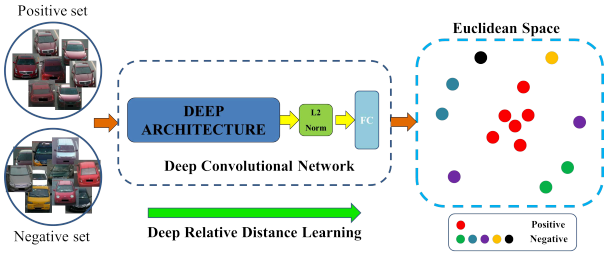

Figure 3. Framework of our model for vehicle re-identification. The deep neural network aims to map the original vehicle images into an Euclidean space that the images of the same vehicle tend to form a cluster while other images tend to locate relatively far away.

2. Related Work

Most previous object identification research targets at either person or human face. Both of them have long been popular topics in computer vision communities and can be described as an 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[23, 9, 14, 13] and vehicle model verification[21]. But being different from our task, all those methods can only reach the vehicle model level instead of identifying whether two vehicles are exactly the same one. Thus, person re-identification is actually the most closely related problem of ours.

Existed approaches of person re-identification mainly rely on handcrafted features like color or texture histograms and then try to model the transformation of person's ap-

pearance across different cameras[5, 11]. Zhao et al.[24] proposed to make use of mid-level features from automatically discovered hierarchical patch cluster trees for view-invariant and discriminative feature extraction.

Recently, deep convolutional network is also introduced into person re-identification problem. Yi et al.[22] applied a "siamese" deep network which has a symmetric structure with two sub-network to learn pair-wise similarity. Ahmed et al. formulate the person re-identification as a binary classification problem[1] and solve it with a specifically designed deep neural network. 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.

The most similar model of our proposed method is the triplet loss deep convolutional network. Connecting deep convolutional network for feature extraction and a special triplet loss has achieved state-of-the-art performance in both person re-identification and face recognition problems[4, 16]. It is assumed that samples of the same identity should be closer from each other compared to samples of different identities. By optimizing the specifically designed triplet loss function, the network will gradually learn a harmonic embedding of each input image in Euclidean space that tends to maximize the relative distance between the matched pair and the mismatched pair. But generating all possible triplets would result in numerous triplets and most of them are too easy to distinguish that would not make any contribution to the loss convergence in training phase. Either offline or online selection in small batch-size data need to be done in advance.

In this paper, we first present a large-scale dataset that contains not only a large number of vehicles captured by real-world surveillance cameras but also multiple images of each vehicle that were captured across different time or cameras. Each vehicle image is attached with an unique id by its license plate. To the best of our knowledge, there are no similar large-scale datasets before which includes multiple images captured by multiple different cameras for each vehicle. We call this dataset as "VehicleID". Then, we propose an end-to-end framework DRDL that are suited for both vehicle retrieval and vehicle re-identification tasks. Notice that two different vehicles(with different license plates) could be almost the same regarding their appearance if they belong to the same model. We aim to capture both the inter-model difference and intra-model difference between different vehicles.

3. Deep Relative Distance Learning

Technically, there are mainly two core components we need to consider for a re-identification problem: a method for feature extraction and a distance metric to compare features across different images. Previous research on re-identification usually focuses either on designing better handcrafted features or building a more comprehensive metric model. But it is mostly empirically determined about which method for feature extraction and which method for distance metric. Can we have an unified model to accomplish both the two tasks? We believe deep convolutional network could be a good answer. In fact, all deep learning approaches for re-identification we mentioned in section 2 are able to extract features and measure the difference at the same time: for each image feed into the network, we get either its embedding in Euclidean space or the similarity estimation with other images directly. The feature extraction component and distance metric component are both contained inside the network. To utilize the advantage of deep convolutional network in vehicle re-identification problem and inspired by the recently proposed triplet loss[16, 4], we propose an enhanced model in this paper named Deep Relative Distance Learning(DRDL). Different with other existed frameworks, we designed a new loss function to accelerate the training convergence and add an extra branch to measure the instance difference between different vehicles while of the same model. Here for concreteness, we first briefly review the triplet loss and then discuss the details of our proposed model.

3.1. Triplet Loss

In a standard triplet loss network, the inputs are a batch of triplet units $\{< x^a, x^p, x_i^n >\}$ where x^a and x^p belong to the same identity while x^a and x^n belong to the different identities. Let $f(x)$ denote the network's feature representation of image x. For a training triplet $< x^a, x^p, x^n >$, the ideal feature representation of them should satisfy the following constraint:

$$\|f(x^a) - f(x^p)\| + \alpha \leq \|f(x^a) - f(x^n)\| \qquad (1)$$

or equally

$$\|f(x^a) - f(x^p)\|^2 + \alpha \leq \|f(x^a) - f(x^n)\|^2 \qquad (2)$$

where α is a predefined constant parameter representing the minimum margin between matched and mismatched pairs. In addition, to avoid the loss function easily exceeding 0, all image features are constrained in a d-dimensional hypersphere $\|f(x)\|_2^2 = 1$. This normalization step is also performed in [4, 16]. Fig. 4 visualizes Eq.(2) in a more intuitive way. Thus, the loss function can be defined as

$$L = \sum^{N} max\{\|f(x^a) - f(x^p)\|_2^2 + \alpha - \|f(x^a) - f(x^n)\|_2^2, 0\} \qquad (3)$$

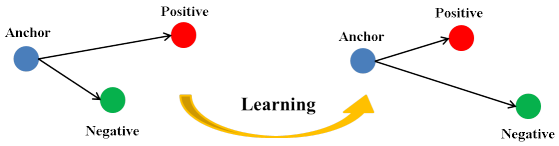

Figure 4. Triplet loss.

However, there exists some special cases that the triplet loss may judge falsely when processing randomly selected triplet units. Given 3 samples that two of them belong to the same identity and the other belongs to a different one, there are two different ways of building triplets as the network's input data. Fig. 5 shows both of them. In the left case, the triplet loss can easily detect the abnormal distance relationship since the intra-class distance is larger than inter-class distance. But the right case is a bit different. The triplet loss is 0 since the distance between anchor and positive point is indeed smaller than the distance between anchor and negative point. Thus, the network will just neglects this triplet during backward propagation.

Figure 5. Two different cases when building triplets.

Moreover, since the backward propagation of triplet loss is actually pulling positive point toward anchor point and pushing negative point toward the opposite direction of anchor point, the loss function is quite sensitive to the selection of anchor point. This means improper anchors can result in great interference in the training stage and lead to a slow convergence. We need a lot more proper triplets to correct it.

3.2. Coupled Clusters Loss

In order to make the training phase more stable and accelerate the convergence speed, we propose a new loss function to replace the triplet loss here: coupled clusters loss(CCL). We also use a deep convolutional network to extract features for each image here. But the triplet input is replaced by two different image sets: one positive set and one negative set. The former set $X^p = \{x_1^p, \cdots, x_{N^p}^p\}$ contains N^p images of the same identity and the other one $X^n = \{x_1^n, \cdots, x_{N^n}^n\}$ contains N^n images of other different identities. It is assumed that samples belong to the same identity should locate around a common center point in the d-dimensional Euclidean space. Thus, samples in the positive set should form a cluster together and samples in the negative set should stay relatively far away. Fig. 7 illustrates the ideal situation.

We first estimate the center point as the mean value of all positive samples

$$c^p = \frac{1}{N^p} \sum_i^{N^p} f(x_i^p) \tag{4}$$

The relative distance relationship is reflected as

$$\begin{gathered} \|f(x_i^p) - c^p\|_2^2 + \alpha \leq \|f(x_j^n) - c^p\|_2^2 \\ \forall 1 \leq i \leq N^p \text{ and } 1 \leq j \leq N^n \end{gathered} \tag{5}$$

The coupled clusters loss function is defined as

$$\begin{gathered} L(W, X^p, X^n) = \\ \sum_i^{N^p} \tfrac{1}{2} max\{0, \|f(x_i^p) - c^p\|_2^2 + \alpha - \|f(x_*^n) - c^p\|_2^2\} \end{gathered} \tag{6}$$

where x_*^n is the nearest negative sample to the center point. If $\|f(x_i^p) - c^p\|_2^2 + \alpha - \|f(x_*^n) - c^p\|_2^2 \leq 0$, the partial derivative of both the positive and negative samples are 0. Otherwise the partial derivative of the positive samples are

$$\frac{\partial L}{\partial f(x_i^p)} = f(x_i^p) - c^p \tag{7}$$

The partial derivative of the nearest negative sample is

$$\frac{\partial L}{\partial f(x_*^n)} = c^p - f(x_*^n) \tag{8}$$

The main idea behind Eq.(5) is absolutely the same as the triplet constraint in Eq.(2) that intra-class distance should be smaller than the inter-class distance. But the way we express it is quite different:

- Distances are measured between samples and a cluster center rather than any randomly selected anchor samples;
- The coupled clusters loss function is defined over multiple samples instead of three.

The first requirement ensures the distances we get and the direction the samples will be moved to(the loss function's partial derivative of each input sample) in backward propagation step are more reliable than the original triplet loss since the randomly selected anchor is replaced by the cluster center. Then, the specifically designed loss function guarantees all positive samples which are not close enough to the center will move closer(samples which are already close enough will be neglected). The selection of the nearest negative sample x_*^n will further prevent the relative distance relationship Eq.(5) being too easily satisfied compared with a randomly selected negative reference.

3.3. Mixed Difference Network Structure

There is a small but quite important difference between identifying a specific vehicle and person. In theory, any two

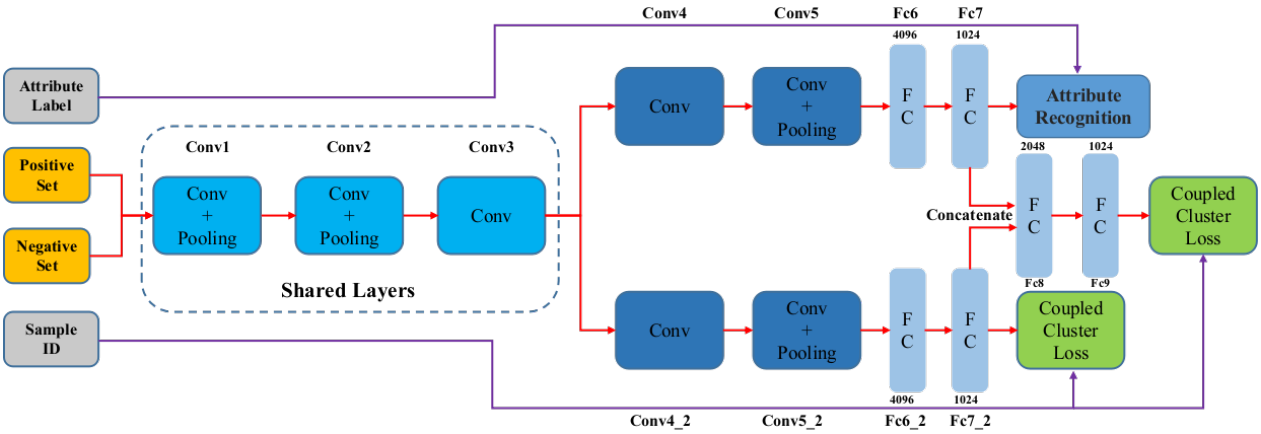

Figure 6. Mixed difference network structure based on *VGG_CNN_M_1024*.

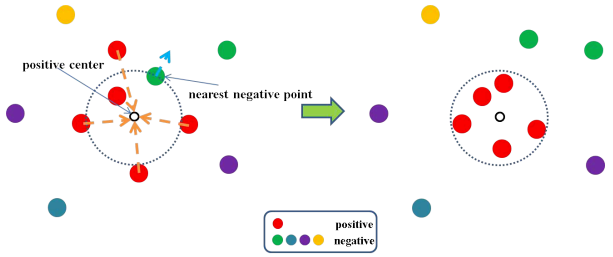

Figure 7. Coupled clusters loss.

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 vehicle model(for instance, they both are Audi A4L 2009). But in real-world scenes, it is still possible to distinguish two vehicles of the same model if some special markers exist which are showed previously in Fig. 2. To deal with this case, the distance measurement between a probe image and multiple candidate vehicles should include two kind of differences: whether they belong to the same vehicle model and whether they are the same vehicle. Since existing models for person re-identification did not really consider this, we propose a new network structure to better measure these two difference here.

The base network structure used in our experiments is *VGG_CNN_M_1024*[2]. It contains 5 convolutional layers and 2 fully-connected layers. The dimension of the network's last fully-connected layer *"fc7"* is 1024.

But since the single branch network structure is not capable of extracting both the vehicle model information and the instance difference between two candidates of the same vehicle model both, we extend the single branch network to a two branches network. Fig. 6 illustrates the detailed structure.

Notice that the last fully connected layer *"fc8"* is a mixed feature of both the vehicle's model information and the feature representation learned from single triplet loss or our coupled clusters loss. The idea behind is quite simple: two vehicle images are definitely different vehicles if they are of different vehicle model and in the other case, i.e. they belong to the same vehicle model, an extra instance difference measurement is needed(The dimension of *"fc8"* is set to 1024 in accordance with the output dimension of standard *VGG_CNN_M_1024* network to eliminate the influence of feature dimensional difference when performing evaluation experiments). *"fc7_2"* in the mixed difference network is just the same as the output feature of a standard *VGG_CNN_M_1024* network while *"fc8"* is an enhanced one suitable for both inter-model difference and intra-model difference metric.

3.4. Training the Network

All the networks in our experiments are trained with the widely-used deep learning framework "Caffe"[8]. Training data consist of multiple positive and negative image sets and the corresponding labels(i.e. ID and vehicle model). Specifically in our experiments, our networks are all fine tuned on *VGG_CNN_M_1024* which is pre-trained with the ImageNet dataset in ILSVRC-2012[15]. We use a momentum of $\mu = 0.9$ and weight decay $\lambda = 2 \times 10^{-4}$. The sizes of both the positive and negative sets are set to 5(images). Batch-size is set to 15 which means we need to feed $15 * (5 + 5) = 150$ images in each training iteration. We start with a base learning rate of $\eta^{(0)} = 0.01$ and then drops by repeatedly multiply 0.7 after every 8000 batch iterations: $\eta^{(i)} = \eta^{(0)} 0.7^{\lfloor i/8000 \rfloor}$. Loss weights of the upper branch(softmax), lower branch(CCL) and the final one(CCL) are 0.5, 0.5, 1.0 respectively.

4. VehicleID Dataset

As mentioned in section 1, the identity information is not available in "CompCars" dataset[21] and we aim to go further than the existed vehicle model recognition task, a more

Figure 8. VehicleID Dataset. Each vehicle have at least 2 images in our dataset. Illumination and viewpoint could both varied a lot in different images. A large number of different vehicles are of the same mode.

suitable large-scale vehicle dataset named "VehicleID"[1] for re-identification task has been carefully collected and organized by us. 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 images. Thus, this dataset could be well suitable for vehicle search related tasks.

In addition, we carefully labeled 10319 vehicles(90196 images in total) of their vehicle model information. But different from the "CompCars" dataset[21], our dataset does not targets 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"). Fig. 9 shows the exact numbers of each vehicle model. As we can see, the most commonly seen vehicle models like "Buick Excelle", "Chevrolet Cruze" and "Volkswagen Lavida" all have more than 200 different vehicles(2000+ images) each, while the most seldom seen vehicle models like "C-Quatre", "Toyota Prado" and "Subaru Forester" have only 1 vehicle of each model.

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. 8 demonstrates some examples.

To better assist different vehicle-related experiments, this dataset is split into two parts for model training and testing. The first part contains 110178 images of 13134 vehicles and 47558 images have been labeled with vehicle model information. The second part contains 111585 images of 13133 vehicles and 42638 images have been labeled with vehicle model information.

[1] Available at http://pkuml.com/resources/pku-vehicleid.html

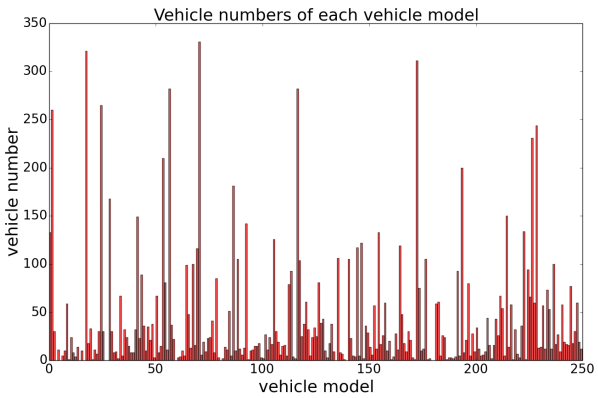

Figure 9. Vehicle numbers of 250 models.

Considering the total number of testing data is still too large compared with ordinary testing data for person re-identification(316 pedestrians in VIPeR dataset[7], 50 in iLIDS dataset[20]), we further extract three subsets(i.e. small, medium and large) ordered by their size from the original testing data for our vehicle retrieval and vehicle re-identification tasks. The quantity distribution of "VehicleID" is demonstrated in Table 1 and Table 2.

Table 1. Data Split For Training And Testing

Image number	Training	Testing
With model label	47558	42638
Without model label	67540	68947
All	110178	111585

Table 2. Test Data Split

	Small	Medium	Large
Number of vehicles	800	1600	2400
Number of images	7332	12995	20038

5. Experiments

Two datasets including "CompCars"[21] and our "VehicleID" dataset are used to evaluate our method. Both of them have a large collection of different vehicle pictures but originally designed for different tasks.

"CompCars" is a recently released large-scale and comprehensive image database, much larger both in scale and diversity compared with other existed vehicle image datasets. There are 214, 345 images(collected from the Internet) of 1, 716 car models in total and the entire dataset has been split into three subsets. Following the pipeline of car model verification[21], we use the Part-I subset which contains 431 car models with a total of 30955 images capturing the entire car for model training and evaluate our method in

vehicle verification task on the Part-III subset which contains 22236 images of 1145 models. Notice that the Part-II subset which contains a list of matched or mismatched pair data is not being utilized when training our model because our network is not designed for pair-wise similarity learning.

We use *VGG_CNN_M_2048*[2] and its mixed difference version in section 3.3 as the feature extractor in all our experiments. Theoretically, other convolutional networks like *GoogleNet*[19], *VGG_ILSVRC_19_layers*[17] can also be embedded in our framework. To make a comprehensive evaluation of our proposed model, we designed 3 different experiments all together: vehicle model verification, vehicle retrieval and vehicle re-identification. The detailed description and the final results are in the following three subsections.

5.1. Vehicle Model Verification

Since vision based vehicle re-identification problem is never deeply explored before and most vehicle search techniques are based on analyzing vehicle model information. We first perform the vehicle model verification task following the pipeline of face verification on "CompCars" dataset to give an overview of our proposed method. This task can be described as: given two vehicle images, we need to verify whether they belong to the same vehicle model. Notice that neither our method nor the standard triplet loss network is designed for vehicle model verification or classification tasks. We did not really expect our model to have a comparable result compared with other mature solutions from face verification problem at first.

Three other methods are introduced to perform the comparison experiments. The experimental results of the first two methods, "Deep Feature+SVM or Joint Bayesian", are referred from Yang's paper[21]. They first utilize a deep convolutional network to train a vehicle model classification model on Part-I data of "CompCars". Then, Joint Bayesian[3] or SVM is applied to train a verification model on Part-II data with the classification network in step 1 as a feature extractor. The third algorithm, *"VGG_CNN_M_1024"* network with triplet loss function is trained with Part-I data of "CompCars" and the corresponding vehicle model labels. Part-II data is not being used since pair-wise data is not suitable for the network's input.

The training process of our method is quite similar to the triplet loss network except that the vehicle make information is also introduced to assist the model training and the standard *"VGG_CNN_M_1024"* is replaced by its mixed difference version. To describe it more specifically, the first branch of the mixed difference network aims to learn the vehicle make information of the input images and the second branch aims to learn the difference between different vehicle model images. The final mixed feature is an enhanced feature representation of the input image. Table 3 presents the final verification accuracy of the above methods.

Table 3. Predict Accuracy of Vehicle Verification Task

Accuracy	Easy	Medium	Hard
GoogleNet+SVM[21]	0.700	0.690	0.659
GoogleNet+Joint Bayesian[21]	0.833	0.824	0.761
VGG+Triplet Loss	/	/	/
Mixed Diff+CCL(Ours)	0.833	0.788	0.703

In this task, the "GoogleNet+Joint Bayesian" framework achieves the best performance on all three testing datasets and our approach ranked the second place overall. The "VGG_CNN_M_1024+Triplet Loss" got no results because its loss function failed to converge during the training phase. Maybe it is just not that easy to form perfect clusters of various different samples simply distinguished by their vehicle model information.

5.2. Vehicle Retrieval

Another closely related problem of vehicle search is the object retrieval task, we evaluate the performance of our proposed DRDL 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. Thus, only three methods are included in this part: "VGG+Triplet Loss", "VGG+Coupled Clusters Loss" and "Mixed Difference VGG+Coupled Clusters Loss". Moreover, as small-scale dataset may affect the deep model's final test accuracy and we aim to fully evaluate the potential power of different models, the entire training data(110178 images in total) in "VehicleID" is being used for model training. In test phase, suppose we have N_i images for vehicle model i, we put $max\{6, N_i - 1\}$ images into the gallery set and the rest into the probe set. After extracting the normalized features using the trained deep convolutional network, the difference between arbitrary two vehicle images is measured directly by their L2 distance. Table 4 illustrates the final results.

Since the training data is absolutely sufficient now, it would not be a difficult problem any more for the standard triplet loss to converge. In all three testing datasets, the mean average precision keeps growing significantly after applying our proposed coupled clusters loss function and the mixed difference network structure compared with the original triplet loss framework. It strongly proves the significant effects of our proposed approach.

Table 4. MAP of Vehicle Retrieval Task

MAP	Small	Medium	Large
VGG+Triplet Loss[4]	0.444	0.391	0.373
VGG+CCL(Ours)	0.492	0.448	0.386
Mixed Diff+CCL(Ours)	0.546	0.481	0.455

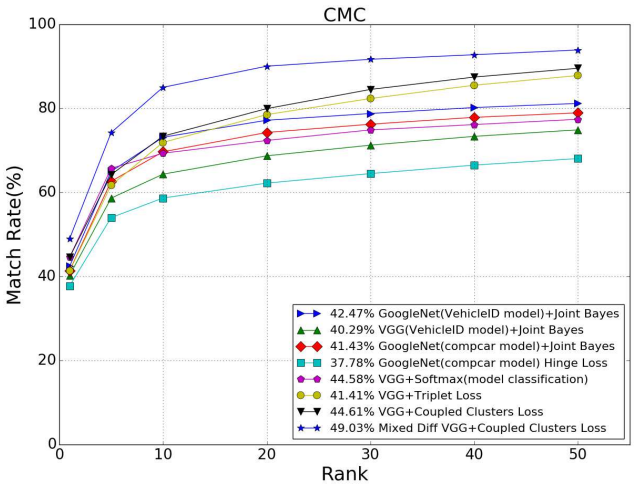

Figure 10. CMC on VehicleID Dataset(gallery size=800).

5.3. Vehicle Re-identification

In this part, we adopted the widely used cumulative match curve(CMC) approach[6] in person re-identification problem and performed the vehicle re-identification experiments on "VehicleID" dataset. For each test dataset split in Table 2, we randomly select one image of each vehicle and put it into the gallery set. Other images are all probe queries. The detailed information of the gallery set and the probe set in each test subset is in Table 5. Following the common method when evaluating model predict accuracy, we repeat it 10 times in testing phase to get the final CMC curve.

Table 5. Gallery and Probe Split for Vehicle ReID Task

Number of images	Small	Medium	Large
Gallery size	800	1600	2400
Probe size	6532	11395	17638

Table 6. Match Rate of Vehicle ReID Task

Match Rate		Small	Medium	Large
VGG+Triplet Loss[4]	top 1	0.404	0.354	0.319
VGG+CCL(Ours)		0.436	0.370	0.329
Mixed Diff+CCL(Ours)		0.490	0.428	0.382
VGG+Triplet Loss[4]	top 5	0.617	0.546	0.503
VGG+CCL(Ours)		0.642	0.571	0.533
Mixed Diff+CCL(Ours)		0.735	0.668	0.616

The detailed match rate from top-1 to top-50 of the various models evaluated on the small-scale test data(gallery size is 800) is illustrated in Fig 10. Considering the number of vehicle models in "CompCars" are far more larger than our "VehicleID" dataset and *"VGG_CNN_M_1024"* is a relative small network for multi-class classification, we also trained a more powerful network, i.e. *"GoogleNet"*, on both datasets. From the results, we can see that when using the same learning model(like the Joint Bayesian), *"GoogleNet"* beats *"VGG_CNN_M_1024"* about 2% in top-1 matching rate and owns significant advantages from top-5 to top-50($6\% - 7\%$). On the other hand, though the various vehicle model information the "CompCars" dataset has, the image type difference(web-nature images and surveillance-nature images) greatly affected the overall predict accuracy. *"GoogleNet"* trained on "VehicleID" beats the one trained on "CompCars" about 3%. The standard "CNN+Triplet Loss" framework works well in this experiment, outperforms all other models except ours. After applying our proposed coupled clusters loss and the mixed difference network structure, the match rate further increased about $4\% - 10\%$ along the CMC curve. Table 6 illustrates the top-1 and top-5 match rate of the best three models on all three test data splits, which reveal the significant advantages of our method again.

6. Conclusions

In this paper, we propose a Deep Relative Distance Learning (DRDL) model to solve an important but not well explored problem: vehicle re-identification. We exploits a two-branch deep convolutional network to map vehicle images into an Euclidean space thus, L2 distance can be directly used for similarity estimation. Compared with existed methods, the specifically designed coupled clusters loss function and the mixed difference network structure achieves a high predict accuracy. Experimental results demonstrate that DRDL achieves promising results and outperforms several state-of-the-art approaches. Although the methods is proposed for vehicle Re-ID, it could work well on vehicle model verification and vehicle retrieval tasks either. For the lack of vehicle re-identification datasets, we present a carefully-labeled large-scale dataset named "VehicleID", which includes multiple images of the same vehicle captured by different real-world cameras. With the nearly 200,000 images of 26,267 vehicles and well-organized identity label, the dataset could easily be used for future research on vehicle re-identification or fine-grained vehicle model recognition tasks.

Acknowledgements

This work is partially supported by the National Basic Research Program of China under grant 2015CB351806, the National Natural Science Foundation of China under contract No. 61425025, No. 61390515, No. 61471042, and No. 61421062, National Key Technology Research and Development Program under contract No. 2014BAK10B02, and Shenzhen Peacock Plan.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3908–3916. IEEE, 2015.

[2] 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.

[3] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012.

[4] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3. Citeseer, 2007.

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision–ECCV 2008*, pages 262–275. Springer, 2008.

[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] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 554–561. IEEE, 2013.

[10] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies*, 17(6):586–606, 2009.

[11] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3610–3617. IEEE, 2013.

[12] W.-H. Lin and D. Tong. Vehicle re-identification with dynamic time windows for vehicle passage time estimation. *Intelligent Transportation Systems, IEEE Transactions on*, 12(4):1057–1063, 2011.

[13] Y.-L. Lin, V. I. Morariu, W. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *Computer Vision–ECCV 2014*, pages 466–480. Springer, 2014.

[14] K. Ramnath, S. N. Sinha, R. Szeliski, and E. Hsiao. Car make and model recognition using 3d curve alignment. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 285–292. IEEE, 2014.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[16] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015.

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

[18] 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.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[20] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *Computer Vision–ECCV 2014*, pages 688–703. Springer, 2014.

[21] 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.

[22] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014.

[23] 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.

[24] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 144–151. IEEE, 2014.