The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning Coarse-to-Fine Structured Feature Embedding for Vehicle Re-Identification

Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, Hanqing Lu
National Laboratory of Pattern Recognition, Institute of Automation
University of Chinese Academy of Sciences, Beijing, China, 100190
{haiyun.guo, chaoyang.zhao, zhiwei.liu, jqwang, luhq}@nlpr.ia.ac.cn

Abstract

Vehicle re-identification (re-ID) is to identify the same vehicle across different cameras. It's a significant but challenging topic, which has received little attention due to the complex intra-class and inter-class variation of vehicle images and the lack of large-scale vehicle re-ID dataset. Previous methods focus on pulling images from different vehicles apart but neglect the discrimination between vehicles from different vehicle models, which is actually quite important to obtain a correct ranking order for vehicle re-ID. In this paper, we learn a structured feature embedding for vehicle re-ID with a novel coarse-to-fine ranking loss to pull images of the same vehicle as close as possible and achieve discrimination between images from different vehicles as well as vehicles from different vehicle models. In the learnt feature space, both intra-class compactness and inter-class distinction are well guaranteed and the Euclidean distance between features directly reflects the semantic similarity of vehicle images. Furthermore, we build so far the largest vehicle re-ID dataset "Vehicle-1M"[1] which involves nearly 1 million images captured in various surveillance scenarios. Experimental results on "Vehicle-1M" and "VehicleID" demonstrate the superiority of our proposed approach.

Introduction

Recent years have witnessed an explosive growing requirement of vehicle re-identification (re-ID) from massive surveillance video in public security field. Similar to pedestrian re-ID, vehicle re-ID is to identify the same vehicle across different cameras. As a unique ID of a vehicle, license plate has been widely used for vehicle re-ID. However, license plate recognition is quite sensitive to image quality, camera view and occlusion. Furthermore, license plate may be removed, altered even faked in some cases, making it unreliable to identify a vehicle simply by its license plate. Therefore, vehicle re-ID by visual appearance is of great practical value in real-world applications.

Different from vehicle model verification, which is to tell whether two vehicles belong to the same vehicle model, vehicle re-ID is more fine-grained and aims to distinguish whether two images contain the identical vehicle. It's a more challenging task and there are few previous attempts purely by visual appearance. The main reasons are two-fold. For one thing, the intra-class and inter-class variation of vehicle images are quite complex, as illustrated in Fig 1. Images of the same vehicle may bear little resemblance to each other, due to illumination change and viewpoint shift. While different vehicles may look quite similar, especially when they belong to the same vehicle model. In this case, we can only differentiate two vehicles based on the special marks, such as the inspection marks and decorations as shown in the circles in Fig 1. Additionally, even when two vehicles are from different vehicle models, it can still be hard to distinguish them since the difference between some vehicle models, such as "BMW-3-2005" and "BMW-3-2009", is very subtle. For another, there is a lack of large-scale dataset for vehicle re-ID. Most existing vehicle datasets, such as *CARS196* (Krause et al. 2013), *CARS333* (Xie et al. 2015) and *CompCars* (Yang et al. 2015), are designed for vehicle model categorization and unfit for vehicle re-ID due to the lack of vehicle ID label. There are only a handful of vehicle re-ID datasets by now, including *VeRi* (Liu et al. 2016b), *VeRi-776* (Liu et al. 2016c) and *VehicleID* (Liu et al. 2016a), which are relatively small and cover limited surveillance scenarios.

Figure 1: The complex intra-class and inter-class variation for vehicle re-ID. The column from left to right contains images of the same vehicle, vehicles of the same vehicle model and vehicles from different vehicle models respectively.

Recently, Convolutional Neural Network (CNN) has achieved state-of-the-art performance in various vision

[1]Available at http://www.nlpr.ia.ac.cn/iva/homepage/jqwang/Vehicle1M.htm.

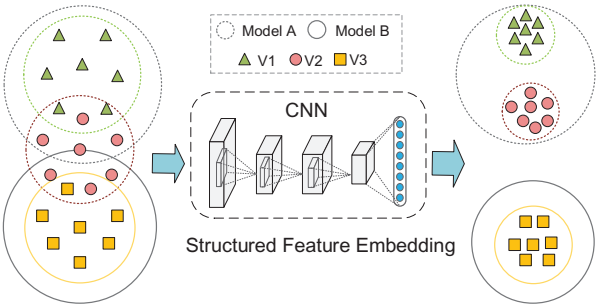

Figure 2: Illustration of the structured deep feature embedding for vehicle re-ID. In the input image space, it's hard to distinguish different vehicles "V1","V2"and "V3". Even different vehicle models "A" and "B" are confusing. With the structured feature embedding, images from the same vehicle are clustered compactly and the discrimination between different vehicles as well as different vehicle models is enhanced.

recognition tasks, such as large-scale image classification (Hu, Shen, and Sun 2017), face recognition (Liu et al. 2017) and person re-ID (Zhao et al. 2017) etc. With the supervision of a carefully designed objective loss, typically contrastive loss (Hadsell, Chopra, and LeCun 2006) or triplet loss (Weinberger and Saul 2009), CNN can learn efficient deep feature embedding to effectively reduce the intra-class variation and enlarge the inter-class variation of input images. The contrastive loss aims to pull matched image pairs closer and require the distance between mismatched ones to be larger than a margin. It can ensure intra-class compactness but the inter-class discrimination is weak. While the triplet loss tries to enforce a distance margin between matched and mismatched image pairs. It can effectively enhance the inter-class discrimination and has gained more popularity than contrastive loss. Lately Liu *et al.* (2016a) adopted an improved version of triplet loss, coupled clusters loss, to learn effective deep features for vehicle re-ID and achieved the best performance so far. Nevertheless, the intra-class compactness is not strong enough to address the large variation between images of the identical vehicle. Besides, they ignored the discrimination between vehicles from different vehicle models, which is actually quite important to obtain a correct ranking order for vehicle re-ID task since images from different vehicle models definitely do not belong to the same vehicle.

In this paper, we propose a novel coarse-to-fine ranking loss to learn a structured deep feature embedding for vehicle re-ID. As shown in Fig 2, the feature embedding can pull images of the identical vehicle together compactly and well discriminate images from different vehicles as well as vehicles from different vehicle models. Thus the Euclidean distance between features in the embedding space can directly reflect the semantic similarity of vehicle images. The coarse-to-fine ranking loss consists of a vehicle model classification loss term, a coarse-grained ranking loss term, a fine-grained ranking loss term and a pairwise loss term, which enables CNN to learn the feature embedding in a coarse-to-fine manner. Firstly, the vehicle model classification loss can cluster vehicles of the same vehicle model together and obtain separability between different vehicle models. The learnt deep feature can well capture the discriminative vehicle model information but the unique vehicle details are eliminated. Besides, the discrimination between different vehicle models is not strong enough. Then we add a coarse-grained ranking loss, which aims to enforce the distance between vehicles of the same vehicle model to be smaller than the distance between those from different vehicle models. It can preserve moderate difference between vehicles of the same vehicle model and enhance the distinction between different vehicle models in the meantime. Next, to reduce the large variation existing in images of the same vehicle and high similarity between vehicles within the same vehicle model, we construct a fine-grained ranking loss which tries to separate images of the same vehicle from those of different ones by a distance margin. Finally, to further restrain the variation between images of the same vehicle, we present a pairwise loss to pull images of the same vehicle as close as possible to each other, even mapped to the same point in the feature space ideally.

In addition, considering the deficiencies of the existing vehicle datasets and the urgent need of large-scale dataset for training deep neural network, we build a large-scale dataset "Vehicle-1M". It's so far the largest vehicle re-ID dataset and covers various real-world surveillance scenarios. Specifically, "Vehicle-1M" includes nearly 1 million images of 55,527 vehicles captured across day and night, from head or rear, by multiple surveillance cameras installed in several cities. Besides, apart from vehicle ID label, each vehicle is attached with one of 400 refined vehicle models, indicating the make, model and year of the vehicle. The difference between vehicle models can be quite small, just like the real-world vehicle re-ID situation. Thus "Vehicle-1M" is suitable for vehicle model categorization, vehicle model verification and vehicle re-ID.

Related Work

Here we briefly review literatures in three categories: vehicle re-ID methods, vehicle re-ID datasets and deep feature embedding methods.

Vehicle Re-ID Methods Though vehicle re-ID has been discussed for many years, most previous works rely on various non-vision based sensors (Kwong et al. 2009; Lin and Tong 2011). Recently, several vision-based vehicle re-ID methods have been proposed. Zapletal *et al.* (2016) solved the vehicle re-ID task by a linear regressor with color histograms and histograms of oriented gradients. However, both the hand-crafted features and the linear model they used are not discriminative enough to address the complex inter-class and intra-class variation of vehicle images. Liu *et al.* (2016b) evaluated hand-crafted features like SIFT and Color Name as well as deep features extracted from CNN models for vehicle re-ID. Experiment results demonstrated that deep features are more discriminative than hand-crafted features. But the CNN model they used for feature extraction was trained for vehicle model classification. The learnt

deep feature could separate vehicles from different vehicle models but the distinction between different vehicles within the same vehicle model was nearly eliminated, thus it was unfit for vehicle re-ID task. Later on, they proposed a progressive vehicle re-ID method "PROVID", which adopted a two-stage search process: coarse-to-fine search in the feature space and near-to-distant search in the real-world surveillance environment. But "PROVID" required extra license plate and spatiotemporal label, which limited its application considerably. Lately, Liu *et al.* (2016a) presented a deep relative distance learning method (DRDL) for vehicle re-ID. They exploited a two-branch CNN supervised by a coupled clusters loss, which tried to separate images from different vehicles from those of the identical vehicle by a distance margin, to learn an Euclidean space where distance can be directly used to measure the similarity of vehicle images. Our approach shares the same spirit with it, but we posed stronger constraints on the compactness between images of the same vehicle and enforce discrimination not only between images from different vehicles but also between vehicles from different vehicle models.

Vehicle Re-ID Datasets Zapletal *et al.* (2016) built the first vehicle re-ID dataset, including 1232 vehicle image pairs. But they were annotated ambiguously, with only people's opinion about whether two images likely to be the same vehicle. The lack of explicit vehicle ID label for each vehicle image made it actually not suitable for vehicle re-ID. Then Liu *et al.* (2016b) proposed "VeRi", which contains over 40,000 images of 619 vehicles and 10 vehicle types. Although "VeRi" involves various viewpoint shift, the illumination change and surveillance scenes are limited. Additionally, it only roughly divides all the vehicles into 10 vehicle types, such as sedan, SUV and truck. Later on, they extended "VeRi" to "VeRi-776" by collecting more vehicle images and adding license plate and spatiotemporal label (Liu et al. 2016c). As "VeRi-776" is mainly an extension of "VeRi", it has the same limitation. Recently, Liu *et al.* (2016a) released a larger dataset "VehicleID", including 221,763 images of 26,267 vehicles and 250 vehicle models. But only 90,196 images are attached with vehicle model label and the vehicle models are not refined enough. Besides, the illumination change is limited. In this paper, we propose so far the largest vehicle re-ID dataset "Vehicle-1M", which is much larger than the existing datasets in the number of image, vehicle and vehicle model. It involves more challenging real-world surveillance scenarios and contains 400 refined vehicle models. The difference between vehicle models can be quite small, just like real-world vehicle re-ID situation. Thus, it's more suitable for evaluating the performance of vehicle re-ID methods.

Deep Feature Embedding Methods Recently, researchers have designed a series of objective losses for deep feature embedding and achieved great performance in face recognition (Sun et al. 2014; Schroff, Kalenichenko, and Philbin 2015; Wen et al. 2016; Liu et al. 2017), object retrieval (Guo et al. 2016; Oh Song et al. 2016), person re-ID (Zhou et al. 2017; Chen et al. 2017) and other vision recognition tasks. Sun *et al.* (2014) jointly utilized classification loss and contrastive loss to learn deep features for face recognition. Classification loss enables CNN to learn separable deep features. But for many vision recognition tasks, the deep features need to be not only separable but also discriminative enough to generalize to other unseen classes. The contrastive loss they used is first proposed in (Hadsell, Chopra, and LeCun 2006) and can ensure intra-class compactness but the inter-class discrimination is still weak. Later schroff *et al.* (2015) adopted triplet loss to learn effective deep face representation. Since triplet loss can effectively enlarge the inter-class variation and reduce the intra-class variation in the meantime, many subsequent works focus on improving triplet loss (Liu et al. 2016a; Oh Song et al. 2016), generalizing triplet loss (Chen et al. 2017) or combing triplet loss with contrastive loss (Zhou et al. 2017) to learn effective deep features for their target tasks. Among them, the most similar to our work is (Chen et al. 2017), which proposed a quadruplet loss consisting of two triplet loss terms and tried to obtain correct ranking order for pairs w.r.t different probe images. The balance of two loss terms is controlled implicitly by two margins. However, the coarse-to-fine ranking loss we present aims to enforce a correct ranking order w.r.t the same probe image, and the coarse-grained and fine-grained ranking loss are balanced explicitly by weighted values. Furthermore, our loss contains a pairwise loss term which can achieve stronger intra-class compactness.

Structured Deep Feature Embedding

In this paper, we aim to exploit the discriminative CNN to learn a structured deep feature embedding $f(x; W)$ for vehicle re-ID. It can map the vehicle image x into an Euclidean space where the distance $D(i, j) = \|f(x_i; W) - f(x_j; W)\|_2^2$ directly reflects the semantic similarity of image pair (x_i, x_j). W denotes the weight parameter used to extract the deep features for vehicle images. Below we propose a coarse-to-fine ranking loss to learn this structured feature embedding in a coarse-to-fine manner.

Coarse-to-fine Ranking Loss Formulation

Let $\{(x_i, v_i, m_i)\}_{i=1}^N$ be the set of training samples in a mini-batch, where N denotes the total number of training samples, v_i and m_i is the vehicle ID and vehicle model label of $i-$th training sample respectively. There are totally V vehicles and M vehicle models, thus $v_i = 1, ..., V$ and $m_i = 1, ..., M$. For a vehicle image pair (x_i, x_j), if $v_i = v_j$, there will be $m_i = m_j$; if $m_i \neq m_j$, there will be $v_i \neq v_j$. Thus we split all the image pairs in the training set into three sets: $\mathcal{P}$, $\mathcal{N}_v$ and $\mathcal{N}_m$. $\mathcal{P} = \{(x_i, x_j) \mid v_i = v_j\}$ contains the image pair of the identical vehicle, $\mathcal{N}_v = \{(x_i, x_j) \mid v_i \neq v_j, m_i = m_j\}$ consists of the image pair from different vehicles but sharing the same vehicle model, $\mathcal{N}_m = \{(x_i, x_j) \mid m_i \neq m_j\}$ includes the image pair from different vehicle models. Our coarse-to-fine ranking loss consists of four loss terms and can be formulated as:

$$L = C + \alpha R_c + \beta R_f + \gamma P, \tag{1}$$

C is the vehicle model classification loss term used to first roughly separate vehicles from different vehicle models. R_c

is the coarse-grained ranking loss term added to enhance the discrimination between vehicles from different vehicle models and enforce moderate distinction between vehicles from the same vehicle model. R_f is the fine-grained ranking loss term utilized to enhance the discrimination between images from different vehicles within the same vehicle model. P is the pairwise loss term adopted to further strengthen the intra-class compactness between images of the identical vehicle. α, β, γ are the weighted values used to balance the four loss terms. Additionally, we constrain all the deep features to live on a $d-$dimensional hypersphere, i.e. $\|f(x)\|_2 = 1$.

Classification loss term Here we implement vehicle model classification loss with cross-entropy loss:

$$C = \frac{1}{N}\sum_{i=1}^{N} -logp_{m_i}, \tag{2}$$

where m_i corresponds to the target class of the input image x_i and p_{m_i} is the predicted probability of image x_i belonging to class m_i. The vehicle model classification loss can well separate vehicles from different vehicle models but the learnt deep features are not discriminative enough to generate to unseen classes. Besides, it clusters vehicles of the same vehicle model so tightly that the distinction between different vehicles is nearly eliminated, which is unfavorable for vehicle re-ID.

Coarse-grained ranking loss term To enhance the discrimination between vehicles from different vehicle models and preserve some variation between vehicles within the same vehicle model, we add the coarse-grained ranking loss:

$$R_c = \frac{1}{\mathcal{Z}_c}\sum_{i=1}^{N}\sum_{(i,j)\in\mathcal{N}_v}\sum_{(i,k)\in\mathcal{N}_m}[D(i,j) - D(i,k) + \mathcal{M}_c]_+, \tag{3}$$

where $\mathcal{Z}_c$ is the normalization factor and $[z]_+$ is the hinge loss $max\{z, 0\}$. This loss tries to enforce a distance margin $\mathcal{M}_c$ between the image pair from the same vehicle model, i.e. (x_i, x_j), and the one from different vehicle models, i.e. (x_i, x_k). To pull vehicles from different models apart as far as possible, for each image x_i, we adaptively select the marginal samples x_k, which are the nearest images from the different vehicle model within the mini-batch. We set $K_1 > 1$ to alleviate the difficulty of CNN convergence.

Fine-grained ranking loss term To further achieve discrimination between images from different vehicles within the same vehicle model, we construct a fine-grained ranking loss term:

$$R_f = \frac{1}{\mathcal{Z}_f}\sum_{i=1}^{N}\sum_{(i,l)\in\mathcal{P}}\sum_{(i,j)\in\mathcal{N}_v}[D(i,l) - D(i,j) + \mathcal{M}_f]_+, \tag{4}$$

where $\mathcal{Z}_f$ is the normalization factor and $\mathcal{M}_f$ is the distance margin used to separate the image pair of the same vehicle, i.e. (x_i, x_l), from the one of different vehicles, i.e. (x_i, x_j). Similarly, we select K_2 marginal samples x_j, which are the nearest images from different vehicle in the mini-batch.

Pairwise loss term To pull images of the same vehicle as close as possible, we add a pairwise loss term to enhance the intra-class compactness. The pair loss is formulated as:

$$P = \frac{1}{\mathcal{Z}_p}\sum_{i=1}^{N}\sum_{(i,l)\in\mathcal{P}} D(i,l), \tag{5}$$

where $\mathcal{Z}_p$ is the normalization factor. This loss can map images of the same vehicle to near points in the feature space, even the same point ideally. Combining the four loss terms, coarse-to-fine ranking loss can achieve high intra-class compactness and discrimination between images from different vehicles as well as vehicles from different vehicle models.

Optimization

We employ the mini-batch based stochastic gradient descent algorithm to optimize the parameter W of CNN. The output loss L for each mini-batch is averaged over L_i for each image x_i. And L_i is formulated as:

$$\begin{aligned} L_i =& C_i + \alpha\sum_{(i,j)\in\mathcal{N}_v}\sum_{(i,k)\in\mathcal{N}_m}[D(i,j) - D(i,k) + \mathcal{M}_c]_+ \\ &+ \beta\sum_{(i,l)\in\mathcal{P}}\sum_{(i,j)\in\mathcal{N}_v}[D(i,l) - D(i,j) + \mathcal{M}_f]_+ \\ &+ \gamma\sum_{(i,l)\in\mathcal{P}} D(i,l), \end{aligned} \tag{6}$$

where C_i denotes the classification loss for image x_i and all the normalization factors are ignored for simplicity. During back-propagation phase, the partial derivative of loss L_i with respect to the deep feature $f(x)$ for each image is given as:

$$\begin{aligned} \frac{\partial L_i}{\partial f(x_i)} =& \frac{\partial C_i}{\partial f(x_i)} + 2\alpha\sum_{(i,j)\in\mathcal{N}_v}\sum_{(i,k)\in\mathcal{N}_m}(f(x_k) - f(x_j))\\ &\mu(J_{ijk}) + 2\beta\sum_{(i,l)\in\mathcal{P}}\sum_{(i,j)\in\mathcal{N}_v}(f(x_j) - f(x_l))\\ &\mu(J_{ijl}) + 2\gamma\sum_{(i,l)\in\mathcal{P}} f(x_i). \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial L_i}{\partial f(x_j)} =& -2\alpha\sum_{(i,j)\in\mathcal{N}_v}\sum_{(i,k)\in\mathcal{N}_m} f(x_j)\mu(J_{ijk})\\ &+ 2\beta\sum_{(i,l)\in\mathcal{P}}\sum_{(i,j)\in\mathcal{N}_v} f(x_j)\mu(J_{ijl}). \end{aligned} \tag{8}$$

$$\frac{\partial L_i}{\partial f(x_k)} = 2\alpha\sum_{(i,j)\in\mathcal{N}_v}\sum_{(i,k)\in\mathcal{N}_m} f(x_k)\mu(J_{ijk}). \tag{9}$$

$$\begin{aligned} \frac{\partial L_i}{\partial f(x_l)} =& -2\beta\sum_{(i,l)\in\mathcal{P}}\sum_{(i,j)\in\mathcal{N}_v} f(x_l)\mu(J_{ijl})\\ &- 2\gamma\sum_{(i,l)\in\mathcal{P}} f(x_l). \end{aligned} \tag{10}$$

where $\mu(z)$ is the indicator function which outputs 1 if $z > 0$ and outputs 0 otherwise. $J_{ijk} = D(i,j) - D(i,k) + \mathcal{M}_c$ and

$J_{ijl} = D(i,l) - D(i,j) + \mathcal{M}_f$. Then the gradient of loss L_i with respect to the weight parameter W is:

$$\frac{\partial L_i}{\partial W} = \sum_{q \in \{i,j,k,l\}} \frac{\partial L_i}{\partial f(x_q)} \frac{\partial f(x_q)}{\partial W}, \tag{11}$$

Vehicle-1M Dataset

Since only a few vehicle re-ID datasets have been proposed by now and the existing ones either cover limited surveillance scenarios or are not large enough for training deep neural network, we carefully build so far the largest vehicle re-ID dataset "Vehicle-1M". Table 1 compares "Vehicle-1M" with other vehicle re-ID datasets in the number of image, vehicle and vehicle model. We collect surveillance video captured during the day and night by multiple real-world surveillance cameras installed in several cities. Then we obtain the compact vehicle images utilizing the vehicle detection algorithm in (Zhu et al. 2016). There are totally 936,051 images of 55,527 vehicles captured from head or rear, across day and night, in multiple surveillance scenes. Apart from vehicle ID label, each image is also annotated with one of 400 refined vehicle models, describing the make, model and year of the vehicle, such as "Benz-S-2011". The difference between vehicle models can be quite small, just like the real-world vehicle re-ID situation, thus this dataset is very suitable for vehicle model categorization, vehicle model verification and vehicle re-ID.

Following the practice in (Liu et al. 2016a), we divide "Vehicle-1M" into the train set, including 844,571 images from 50,000 vehicles, and the test set, which contains the remaining 91,480 images from 5,527 vehicles. Furthermore, inspired by (Liu et al. 2016a), which evaluated methods on small (800 vehicles), medium (1600 vehicles) and large (2400 vehicles) test set separately to test the scalability, we also extract a small, medium and large test set from the original test set, denoted as "Small", "Medium" and "Large" respectively. Specifically, the small test set contains 1,000 vehicles of 16,123 images, the medium covers 2,000 vehicles of 32,539 images and the large one includes 3,000 vehicles of 49,259 images.

Dataset	Image	Vehicle	Model
VeRi	40,000	619	10
VeRi-776	49,360	776	10
VehicleID	221,763	26,267	250
Vehicle-1M	936,051	55,527	400

Table 1: Comparison of different vehicle re-ID datasets.

Experiments

Here we first describe our experiment setup briefly.

Datasets We evaluate our approach on "Vehicle-1M" and "VehicleID", the top two largest vehicle re-ID datasets by now. Each of them has at least two images captured from different cameras for a vehicle.

Evaluation Protocol To conduct vehicle re-ID experiment, we split the test set into probe set and gallery set. The results both on "Vehicle-1M" and "VehicleID" are measured by Cumulative Matching Characteristic (CMC) curve and Mean Average Precision (MAP). CMC is an estimation of finding the correct match in the top K returned results. To calculate this criteria on "Vehicle-1M", we randomly select one image of each vehicle and put it into the gallery set. Other images are all probe queries. MAP is a comprehensive index which considers both the precision and recall of the results. To compute MAP on "Vehicle-1M", we adopt an opposite way. We randomly select one image for each vehicle and put it into the probe set. The remaining forms the gallery set. Final CMC and MAP is averaged over ten repeats of the random splitting process. As for CMC and MAP calculation on "VehicleID", we refer to (Liu et al. 2016a).

Parameter Setting The proposed deep feature embedding learning framework is compatible with any CNN architecture, such as inception-like model (Szegedy et al. 2016) and ResNet (He et al. 2016). But for fair comparison, here we adopt GoogleNet and initialize it with the model pre-trained on CompCars for vehicle model classification. In test phase, all the deep features are extracted from the "pool5" layer of GoogLeNet. For all the experiments, we fix $\mathcal{M}_c = \mathcal{M}_f = 0.2$ following the practice of (Guo et al. 2016) and set $K_1 = 10, K_2 = 3$ when the mini-batch size is 150. As for α, β and γ, they are firstly used to balance the scale of the four loss terms. From the output loss value on the validation set we found the classification loss value was about 1000 times larger than the fine-grained ranking loss, 100 times larger than the coarse-grained ranking loss and 10 times larger than the pair loss at the beginning of training. The reason are two folds. On the one hand, the feature is firstly normalized before calculating the latter three loss terms for training stability. On the other hand, for the initial feature embedding, the intra-class difference is relatively large and the discrimination between vehicle models is greater than that between vehicles. Thus, to balance the loss value scale, we set α to 100, β to 1000 and γ to 10. Since the fine-grained ranking loss directly pulls images of the same vehicle closer and pushes images from different vehicles far away, which is quite compatible with vehicle re-ID task, we deduce it plays a dominant role for vehicle re-ID. Then we fix β and vary the value of α and γ to adjust the importance of corresponding loss. However, variation from 50 to 500 for α and from 5 to 50 for γ only brings slight performance change on both datasets. And when we set α to 100 and γ to 10, we achieved the best re-ID performance. So we adopt this group of parameter value for the following experiments.

Comparison Methods Considering deep features demonstrate to be more discriminative than traditional hand-crafted features for vehicle re-ID (Liu et al. 2016b), we mainly compare our approach to the deep learning based methods. One is the method proposed in (Liu et al. 2016b), which directly extracted deep features from the GoogleNet pre-trained on CompCars. We denote it as "GoogleNet" and adopt it as our baseline method. The other is "DRDL" (Liu et al. 2016a) which utilized a two-branch CNN and a coupled clusters loss to learn deep features for vehicle re-ID and achieved the best performance on "VehicleID". Furthermore, we design a series of methods to verify the

effectiveness of each loss term in the coarse-to-fine ranking loss. Since the fine-grained ranking loss plays a dominant role for vehicle re-ID, we use it as the base loss term, denoted as "FGR", and add other loss terms gradually. "C+FGR" represents jointly adopting the coarse-grained and fine-grained ranking loss to train the CNN. Then we add the pair loss term to "C+FGR" and obtain "C+FGR+P". "C2F-Rank" is the final coarse-to-fine ranking loss we proposed.

Experimental Results on VehicleID

Table 2 illustrates the top 1 and top 5 match rate of our proposed and other comparison methods for vehicle re-ID on "VehicleID". "GoogLeNet" achieved relatively low performance, showing that deep features extracted from CNN model trained for vehicle model classification are unfit for vehicle re-ID task. Because the distinction between different vehicles within the same vehicle model is reduced greatly. "DRDL" jointly used coupled clusters loss and vehicle model classification loss for CNN training. Since only half of the images in "VehicleID" have vehicle model label, our methods, including "FGR", "C+FGR", "C+FGR+P" and "C2F-Rank", just used the training images that have both the vehicle ID and vehicle model label for CNN training. While "DRDL" utilized all the training images. Nevertheless, our method, just with the base loss "FGR", can outperform "DRDL" on nearly all the evaluation indexes. In addition, "FGR" is more compatible with the vehicle re-ID task than vehicle model classification loss thus significantly improves the re-ID performance compared with "GoogLeNet". "C+FGR" improves the performance by a large margin, namely, 7.6% on average. The great performance increase demonstrates that the discrimination between vehicles from different vehicle models indeed plays an important role for vehicle re-ID. With the join of the pairwise loss term, "C+FGR+P" further promotes the performance by around 1%, verifying the effectiveness of enhancement on intra-class compactness for vehicle re-ID. With the vehicle model classification loss further enlarging the variation between different vehicle models, "C2F-Rank" makes further improvements on vehicle re-ID performance and outperforms the state-of-the-art method "DRDL" by 12.7%. What's more, experiments show that the convergence speed using the pairwise or ranking based constraints is relatively slow and the join of classification loss can effectively alleviate this situation.

Figure 3 shows the detailed match rate results from top 1 to top 50 on the small test set of "VehicleID". We can see that our methods consistently beat "GoogLeNet" by a large margin. Additionally, the join of each loss term also shows a consistent performance improvement, verifying the effectiveness of each loss term again. Table 3 demonstrates the MAP results of all the methods on "VehicleID". The inferior performance of "FGR" compared with "DRDL" is because that the former uses fewer training images. "C+FGR" outperforms "FGR" by about 8%, proving again that the coarse-grained ranking loss plays an important role for vehicle re-ID."C+FGR+P" further increases the performance with 1% around, verifying the effectiveness of the pairwise loss. "C2F-Rank" promotes the performance by 3%, demonstrating the effectiveness of the vehicle model classification loss for vehicle re-ID. At last, "C2F-Rank" outperforms "DRDL" by 9.3%, showing the advantage of our proposed method for vehicle re-ID task again.

Match Rate		Small	Medium	Large
GoogLeNet	top 1	0.464	0.425	0.381
DRDL		0.49	0.428	0.382
FGR		0.502	0.454	0.408
C+FGR		0.589	0.536	0.489
C+FGR+P		0.598	0.540	0.499
C2F-Rank		**0.611**	**0.562**	**0.514**
GoogLeNet	top 5	0.622	0.589	0.554
DRDL		0.735	0.668	0.616
FGR		0.724	0.676	0.630
C+FGR		0.793	0.748	0.699
C+FGR+P		0.811	0.748	0.709
C2F-Rank		**0.817**	**0.762**	**0.722**

Table 2: Match rate of vehicle re-ID task on "VehicleID".

Approach	Small	Medium	Large
GoogLeNet	0.462	0.440	0.381
DRDL	0.546	0.481	0.455
FGR	0.531	0.486	0.424
C+FGR	0.609	0.570	0.502
C+FGR+P	0.625	0.577	0.512
C2F-Rank	**0.635**	**0.600**	**0.530**

Table 3: MAP of vehicle re-ID task on "VehicleID".

Experimental Results on Vehicle-1M

Table 4 and 5 shows the match rate and MAP results on "Vehicle-1M" respectively. Similar to the results on "VehicleID", "GoogLeNet" used by (Liu et al. 2016b) performs relatively low and our methods beat this baseline method by a large margin. Specifically, the fine-grained ranking loss term "FGR" alone outperforms "GoogLeNet" by about 8% and 5%, on match rate and MAP respectively, and the join of each loss term promotes this performance gain consistently. "C+FGR" increases the performance of "FGR" by 1.9% and 2.7%, which is bigger than the promotion brought by the join of other loss terms. Additionally, the pairwise loss makes a further performance improvement of around 0.5% and 1.4%, showing its effectiveness. Finally, the join of vehicle model classification loss brings a performance gain of 0.4% and 0.8% or so, verifying again that separating different vehicle models benefits the vehicle re-ID task. The detailed match rate results from top 1 to top 50 on the small test set of "Vehicle-1M" is illustrated in Figure 3. It can be seen that our methods demonstrate a great advantage over "GoogLeNet" consistently.

In summary, the join of each loss consistently improves the re-ID performance, verifying the effectiveness of them. "C2F-Rank" outperforms the state-of-the-art methods by a large margin, showing the superiority of our coarse-to-

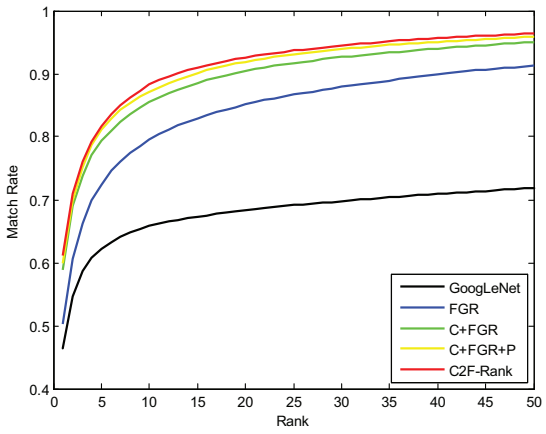

Figure 3: CMC on the "Small" test set of "VehicleID"

Match Rate		Small	Medium	Large
GoogLeNet	top 1	0.545	0.508	0.409
FGR		0.634	0.593	0.500
C+FGR		0.659	0.613	0.519
C+FGR+P		0.667	0.616	0.525
C2F-Rank		**0.671**	**0.620**	**0.528**
GoogLeNet	top 5	0.609	0.586	0.505
FGR		0.683	0.649	0.574
C+FGR		0.691	0.665	0.591
C+FGR+P		0.701	0.667	0.596
C2F-Rank		**0.703**	**0.671**	**0.601**

Table 4: Match rate of vehicle re-ID task on "Vehicle-1M".

Approach	Small	Medium	Large
GoogLeNet	0.682	0.600	0.547
FGR	0.825	0.752	0.692
C+FGR	0.852	0.776	0.724
C+FGR+P	0.866	0.790	0.737
C2F-Rank	**0.871**	**0.798**	**0.747**

Table 5: MAP of vehicle re-ID task on "Vehicle-1M".

fine ranking loss. Furthermore, we can find that the coarse-grained ranking loss makes the most significant performance gain, demonstrating the discrimination between different vehicle models indeed plays a critical role for vehicle re-ID.

Conclusion

In this paper, we propose a novel coarse-to-fine ranking loss to learn a structured deep feature embedding for vehicle re-ID task. The coarse-to-fine ranking loss consists of a vehicle model classification loss term, a coarse-grained ranking loss term, a fine-grained ranking loss term and a pairwise loss term. It aims to pull images of the identical vehicle as close as possible and gain discrimination between images from different vehicles as well as vehicles from different vehicle models in a coarse-to-fine manner. With the supervision of it, CNN can learn a deep feature embedding space where the Euclidean distance between feature points directly reflects the semantic similarity of vehicle images. Besides, we present so far the largest vehicle re-ID dataset "Vehicle-1M", which contains nearly 1 million images from 55,527 vehicles and 400 vehicle models. Our approach achieves the state-of-the-art performance both on "Vehicle-1M" and "VehicleID", showing its advantage for vehicle re-ID task.

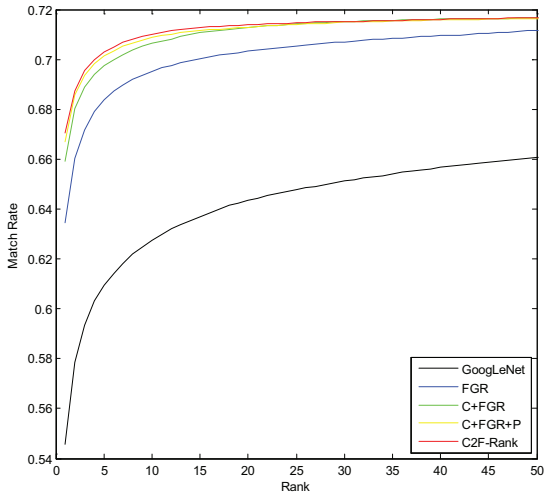

Figure 4: CMC on the "Small" test set of "Vehicle-1M"

Acknowledgments

This work was supported by National Natural Science Foundation of China 61772527.

References

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*.

Guo, H.; Wang, J.; Gao, Y.; Li, J.; and Lu, H. 2016. Multi-view 3d object retrieval with deep embedding network. *IEEE Transactions on Image Processing* 25(12):5526–5537.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, 1735–1742. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 554–561.

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

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

Liu, H.; Tian, Y.; Yang, Y.; Pang, L.; and Huang, T. 2016a. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2167–2175.

Liu, X.; Liu, W.; Ma, H.; and Fu, H. 2016b. Large-scale vehicle re-identification in urban surveillance videos. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, 1–6. IEEE.

Liu, X.; Liu, W.; Mei, T.; and Ma, H. 2016c. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European Conference on Computer Vision*, 869–884. Springer.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. *arXiv preprint arXiv:1704.08063*.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4004–4012.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, 1988–1996.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition*, 2818–2826.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10(Feb):207–244.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 499–515. Springer.

Xie, S.; Yang, T.; Wang, X.; and Lin, Y. 2015. Hyper-class augmented and regularized deep learning for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2645–2654.

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

Zapletal, D., and Herout, A. 2016. Vehicle re-identification for automatic video traffic surveillance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 25–31.

Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; and Tang, X. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, S.; Wang, J.; Wang, J.; Gong, Y.; and Zheng, N. 2017. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, Y.; Wang, J.; Zhao, C.; Guo, H.; and Lu, H. 2016. scale-adaptive deconvolutional regression network for pedestrain detection. In *ACCV*.