# Group-Sensitive Triplet Embedding for Vehicle Reidentification

Yan Bai, *Student Member, IEEE*, Yihang Lou, *Student Member, IEEE*, Feng Gao, *Student Member, IEEE*, Shiqi Wang, *Member, IEEE*, Yuwei Wu, and Ling-Yu Duan, *Member, IEEE*

***Abstract*—The widespread use of surveillance cameras toward smart and safe cities poses the critical but challenging problem of vehicle reidentification (Re-ID). The state-of-the-art research work performed vehicle Re-ID relying on deep metric learning with a triplet network. However, most existing methods basically ignore the impact of intraclass variance-incorporated embedding on the performance of vehicle reidentification, in which robust fine-grained features for large-scale vehicle Re-ID have not been fully studied. In this paper, we propose a deep metric learning method, group-sensitive-triplet embedding (GS-TRE), to recognize and retrieve vehicles, in which intraclass variance is elegantly modeled by incorporating an intermediate representation "group" between samples and each individual vehicle in the triplet network learning. To capture the intraclass variance attributes of each individual vehicle, we utilize an online grouping method to partition samples within each vehicle ID into a few groups, and build up the triplet samples at multiple granularities across different vehicle IDs as well as different groups within the same vehicle ID to learn fine-grained features. In particular, we construct a large-scale vehicle database "PKU-Vehicle," consisting of 10 million vehicle images captured by different surveillance cameras in several cities, to evaluate the vehicle Re-ID performance in real-world video surveillance applications. Extensive experiments over benchmark datasets VehicleID, VeRI, and CompCar have shown that the proposed GS-TRE significantly outperforms the state-of-the-art approaches for vehicle Re-ID.**

***Index Terms*—Vehicle re-identification, metric learning, intra-class variance, embedding, retrieval, surveillance.**

Manuscript received August 10, 2017; revised November 25, 2017; accepted December 27, 2017. Date of publication January 23, 2018; date of current version August 14, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants U1611461, 61661146005, and 61390515, and in part by the National Key Research and Development Program of China under Grant 2016YFB1001501. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhu Li. *(Corresponding author: Ling-Yu Duan.)*

Y. Bai and Y. Lou are with the Institute of Digital Media, Peking University, Beijing 100871 China, and also with the School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China (e-mail: yanbai@pku.edu.cn; yihang@pku.edu.cn).

F. Gao and L. Duan are with the Institute of Digital Media, Peking University, Beijing 100871, China (e-mail: gaof@pku.edu.cn; lingyu@pku.edu.cn).

S. Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: shiqwang@cityu.edu.hk).

Y. Wu is with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wuyuwei@bit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2796240

## I. Introduction

TOWARDS the major strategic needs of the social public security, how to address the grand challenge on video big data is an emerging research area. Cross-view correlation and recognition of objects and events in images/videos big surveillance data is becoming a crucial but challenging research problem. In this work, we focus on the large-scale recognition and retrieval of vehicles in images, which is expected to facilitate the spatial-temporal object recognition and behavior analysis in wide video surveillance networks. Vehicle re-identification (Re-ID) aims to quickly search, locate and track the target vehicles across surveillance camera networks, which plays key roles in maintaining social public security and serves as a core module in the large-scale vehicle recognition, intelligent transportation, surveillance video analytic platforms [1]–[6]. Vehicle Re-ID refers to the problem of identifying the same vehicle in a large scale vehicle database given a probe vehicle image. In particular, vehicle re-identification can be regarded as a fine-grained recognition task [7]–[10] that aims at recognizing the subordinate category of a given class. A typical example is on the fine-grained recognition of a specific vehicle model, such as "Buick Regal 2011 model." However, the granularity of vehicle re-identification task is much finer since the ideal target is to search a specific vehicle rather than a model, in which the image instances of the same vehicle are formed as a separate category. As illustrated in Fig. 1, given two vehicle images of "Buick Regal 2011 model," they should be assigned to different classes with different IDs although they come from the same vehicle model. Hence, discriminative visual features that are capable of representing the subtle characteristic differences, such as specific marks like annual inspection, tissue boxes, ornaments, etc, are required.

The straightforward vehicle re-identification approaches resort to robust license plate recognition [11]–[13], as license plate provides the unique identity information of vehicles. However, license plate recognition often fails in unconstrained surveillance environments. On one hand, various viewpoints, illuminations and imaging resolutions may significantly degrade the license plate recognition accuracy. On the other hand, there exist many hard cases where the license plates of problematic vehicles are actually occluded, removed, or even deliberately faked. To alleviate the limitation of license plate recognition methods, we focus on the effective matching and retrieval of visual features for vehicle re-identification based on discriminative

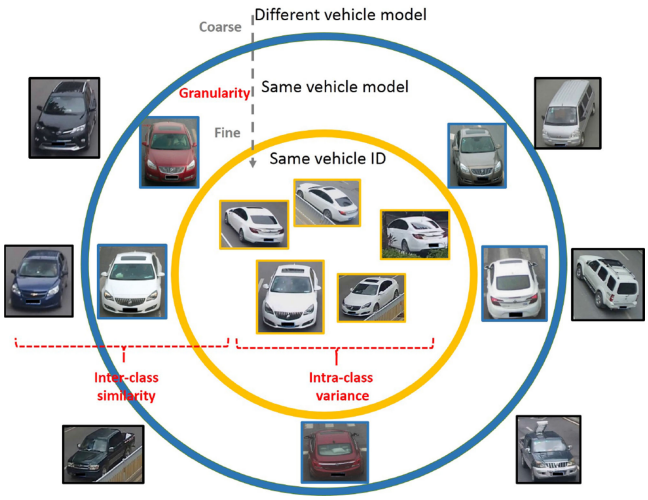



Fig. 1. Illustrations of the coarse-to-fine concept of vehicle recognition at different granularity. The vehicle model recognition aims at recognizing vehicles with a specific vehicle model, while vehicle re-identification is to search out the same vehicle as the query vehicle. The intra-class variance and inter-class similarity present the very natural challenges to vehicle re-identification.

visual appearance, which is crucial for the full-fledged vehicle re-identification systems.

To this end, robust vehicle re-identification has to tackle the issue of extracting discriminative features to distinguish different vehicles. Basically, there exist three challenges. (1) The captured image samples of a vehicle from different cameras may produce large variance in appearance, as shown in Fig. 2(a). This is regarded as intra-class variance, and robust feature representation is crucial. (2) Different vehicles may exhibit fairly similar visual appearance, especially when those different vehicle IDs come from the same model. To tackle the problem of considerable inter-class similarity, the visual features are required to capture and represent the subtle characteristic differences, as shown in Fig. 2(b). (3) In real-world scenarios, the city-scale surveillance systems usually involves millions vehicles while the available well-annotated training images are very limited. A desired approach should be capable of scaling up to deal with large-scale datasets through learning effective feature representation over limited data.

Recently, deep metric learning has achieved great success in various tasks, such as face recognition [14]–[17], visual search [18]–[21], person/vehicle re-identification [22]–[26], fine-grained image recognition [15], [27], [28]. Deep metric learning in these various tasks aims to learn and strengthen feature space embedding that pulls the similar images closer and pushes the dissimilar images far away. In this work, the embedding space is implemented by a deep convolutional neural network optimized by a triplet loss function. To adequately address the above three issues for vehicle re-identification, we propose a group sensitive triplet embedding (GS-TRE) method. Specifically, the deep metric learning is employed to learn a feature embedding space by optimizing the similarity distances between sample features. By introducing an intermediate representation "group" between samples and vehicle IDs, GS-TRE attempts to build up a sort of "similar attribute, closer distance" feature embedding. In particular, for the high intra-class variance, the grouping process is adopted to incorporate the intra-class variance into the metric learning. Given a vehicle ID, samples of the vehicle are clustered into a set of groups, such that in each group the samples have similar attributes. Regarding the inter-class similarity, we treat each vehicle ID as a separate class and jointly optimize the multi-task objective of classification and metric learning, with an aim of constraining the samples of the same vehicle IDs together and pushing the samples of different vehicles away for feature level discrimination. Moreover, the scalability in large-scale re-identification can be resolved by deriving discriminative features via a deep network trained from a small scale training set, and subsequently performing retrieval task in large-scale database instead of classification so as to avoid the curse of excessively large number of classes. Extensive experiments over benchmark vehicle datasets show that the proposed GS-TRE method significantly outperforms the state-of-the-art fine-grained visual recognition approaches.

In summary, the contributions of this paper are three-fold.

- First, we propose a group sensitive triplet embedding approach to modeling the inter-class dissimilarity as well as the intra-class invariance in triplet network learning. The GS-TRE can significantly mitigate the negative impact of inter-class similarity and intra-class variance on the fine-grained recognition, which has been well demonstrated in vehicle re-identification.
- Second, we propose to leverage multi-task learning to generate discriminative feature representation by the joint optimization of group sensitive triplet loss and softmax loss, which can be well applied to accomplish large scale vehicle re-identification towards real applications.
- Third, we construct a large-scale vehicle dataset "PKU-Vehicle" containing 10 millions vehicle images, which are collected from different real-world surveillance cameras in several cities. This dataset may contribute to the comprehensive evaluation of the vehicle re-identification methods, and is expected to push forward the research on fine-grained object recognition. The dataset including fine-grained features is available at http://59.110.216.11/html/

The remainder of this paper is organized as follows: Section II reviews the relevant works, and Section III gives the problem statement of vehicle re-identification from the perspective of metric learning. We introduce the proposed group sensitive triplet embedding in Section IV. Extensive experimental results are presented and analyzed in Section V, and finally the paper is concluded in Section VI.

## II. Related Work

As an emerging research topic, vehicle re-identification has attracted great efforts [3]–[5], [24], [29]–[32]. In this section, we will review the relevant works from three aspects: vehicle re-identification, fine-grained recognition, and deep metric learning.

*Vehicle Re-Identification:* In recent years, the success of Convolutional Neural Network (CNN) [33]–[35] has greatly facilitated research topics on vehicle recognition, such as vehicle


(a) (b)

Fig. 2. Examples of illustrating the intra-class variance and inter-class similarity. (a) The same vehicle captured by different cameras produces the significant intra-class variance due to the different shotting angles or illuminations. (b) Different vehicles of the same model exhibits strong inter-class similarity, in which their subtle characteristic regions marked by red circles are useful to generate discriminative features for fine-grained vehicle recognition.

classification [1], [36], verification [1], [3], and attributes prediction [37], [38]. For analyzing traffic surveillance big video data, high performance vehicle re-identification is becoming a challenging topic. Many vehicle re-identification methods are proposed to retrieve vehicles by the characteristics and attributes of vehicles, such as license plate identification, spatial-temporal property and color. Feris *et al.* [39] proposed a vehicle detection and retrieval system to identify the attributes and colors of vehicles, and performed further retrieval based on the recognized vehicle attributes. Liu *et al.* [24] proposed a vehicle re-identification system to fulfill the coarse-to-fine vehicle search in the feature space, followed by context assisted search in the real-world spatial-temporal environment. Different from the above methods, some research works focus on hybrid features to enhance the recognition of vehicle characteristics. For example, Cormier *et al.* [29] presented a descriptor that combines local binary patterns and local variance, to solve the problem of low resolution vehicle re-identification. Liu *et al.* [3] introduced a mixed difference network for vehicle re-identification, in which the vehicle model features and the metric learning feature are both incorporated into a single network. Despite of the above-mentioned progress on vehicle re-identification, the impact of intra-class variance and inter-class similarity have not been well investigated, which can significantly influence vehicle recognition performance.

*Fine-grained visual recognition:* As mentioned before, vehicle re-identification is a typical example of fine-grained recognition. There are two typical topics in fine-grained vehicle recognition, i.e., part-based model and representation learning model. Many methods [40]–[43] employ part-localization and alignment to extract the features of key parts of the objects and perform detailed comparison on parts. Xiao *et al.* [44] introduced reinforcement learning to adaptively find discriminative regions in fine-grained domains in a weakly-supervised way. Moreover, Zhao *et al.* also came up with a diversified visual attention network to relieve the dependency strongly supervised information for learning to localize key regions. In practice, the dramatically variant shooting angles may result in significantly different visible parts. Hence, several representative works prefer the representation learning approaches instead. Lin *et al.* [45] proposed a bilinear architecture to obtain the local pairwise features where the output features of two separate networks are fused in an invariant manner. Krause *et al.* [46] leveraged noisy data from the web and adopted simple but generic representation learning methods to achieve the state-of-the-art results on several fine-grained benchmarks. Similarly, our method also focuses on representation learning, which emphasizes the optimization of distance metric of samples from the perspective of incorporating the modeling of sample distribution into metric learning. The key idea is to leverage the intra-class structure to model a so-called group sensitive feature distribution, which is able to enhance the fine-grained feature representation.

*Deep Metric Learning:* The inter-class similarity and intra-class variance relate to two basic challenges in feature learning. To resolve these issues, many promising methods [16], [47], [48] leverage deep networks to learn a feature embedding space to maximize inter-class distances and minimize the intra-class distances simultaneously. In particular, a sort of triplet constraint in [47] was introduced to learn a feature embedding based on the principle "the samples belonging to the same vehicle ID are closer than those samples belonging to different IDs." Such triplet constraint has been widely used in pedestrian

re-identification [49]–[52] and face recognition [16] tasks. Based on triplet, a quadruplet network is also proposed by Chen *et al.* [50] to improve the generalization capability of feature representation. In [51], Yang *et al.* leveraged privileged information and unlabeled samples as auxiliary data to construct discriminant metric. In [53], Zhang *et al.* proposed to employ multiple labels to inject hierarchical inter-class relationship (different models, brands, manufactured years, etc) as prior knowledge into learning feature representation, while the effects of intra-class variance in feature distribution are not investigated. Lin *et al.* [54] utilized bipartite-graph labels to model rich inter-class relationships based on multiple sub-category components, which can be elegantly incorporated into convolutional neural network. Wen *et al.* [15] proposed to learn an optimal center for deep features of each class and penalize the distances between the deep features and their corresponding class centers. Moreover, some related works devoted to bring the semantic knowledge to metric learning. Cui *et al.* [55] designed a general knowledge graph to capture the relations of concepts in image representation, then a regularized regression model is leveraged to jointly optmize the image representation learning and graph embedding. Li *et al.* [56] explored how to utilize the user-provided tags to learn a distance metric, which can reflect the semantic information and improve the performance of tag-based image retrieval.

Most of efforts are devoted to optimizing the inter-class distance, while the constraints of local structure of feature space within a class are seldomly studied, which is useful for dealing with large intra-class variance. Accordingly, our approach aims to impose the local structure constraints at the fine granularity within a class into deep metric learning, which is shown to be effective in generating discriminative features.

## III. Problem Statement

The vehicle images acquired from urban surveillance camera networks pose dramatic appearance changes from different angles, occlusions, lighting illuminations and cluttered backgrounds. In particular, as shooting angles or backgrounds in traffic surveillance scenes are diverse but still limited, the inherent appearance variance within each vehicle ID needs proper modeling, which is expected to impact the performance of feature matching of between different vehicles. Therefore, we attempt to group vehicle images to represent the intra-class variance (e.g., angle, color, background), and thereby form a group sensitive structure, in which the vehicle images of each specific group are supposed to share similar attributes. As such, the intra-class variance can be well modeled, which is useful to discriminate the subtle visual appearance differences between vehicles.

Moreover, another critical issue of re-identifying vehicles arises from the big and fast growing scale of vehicles. The number of vehicles in a typical city-scale surveillance system usually reaches up to millions scale. It is infeasible to develop million-scale classifiers to realize vehicle re-identification from the classification point of view. Moreover, it is difficult to collect large-scale well-annotated vehicle datasets. For example, the VehicleID dataset [6], which is the largest vehicle re-identification benchmark dataset to the best of our knowledge, contains 26,267 vehicle IDs. To deal with large-scale vehicle re-identification, we resort to the retrieval solution. Then the remaining issue is to develop discriminative features for representing vehicles at a fine granularity.

Here, we propose to structure the image samples for each vehicle IDs. Let $S^p$ denote a set of samples of a specific vehicle ID $p$ and $S^n$ represents the samples of other vehicle IDs ($p \neq n$). Assume that the instances of each vehicle are divided into $G$ groups, we have $S^{p,g}$ ($g \in \{1, 2, ..., G\}$) to denote a set of instances in group $g$ for the vehicle $p$. Clearly, multiple distinct groups within each vehicle ID are expected to represent intra-class variance. Our aim is to model intra-class structure in each vehicle's feature distribution, and then minimize the distances of samples in the same group for each vehicle ID while keeping the samples apart away from different vehicle IDs with a minimum margin $\alpha$. The optimization objective can be formulated as follows:

$$\begin{aligned} &\min_{M} \sum_{g=1}^{G} \sum_{x_i, x_j \in S^{p,g}} \|x_i - x_j\|_M^2 \\ &\text{s.t.} \sum_{x_i \in S^p, x_n \in S^n} \|x_i - x_n\|_M^2 \geq \alpha \\ &M \succeq 0, \end{aligned} \tag{1}$$

where $x_i$ and $x_j$ denote the samples from the vehicle $p$ falling into the same group $g$, and $x_n$ denotes other vehicles. $M$ is a metric matrix, and $\alpha$ is the minimum margin constraint under $M$ between the samples from different vehicles. In this work deep metric learning is applied to model the intra-class variance in feature space to generate robust and discriminative feature representation.

## IV. GSTE Approach

With the prior of the intra-class variance attributes, a group-level finer representation within each vehicle ID can be characterized and the intra-class variance is presented by a set of groups. To mitigate the negative effects of the intra-class variance and inter-class similarity, GSTE leverages inter-class triplet embedding as well as intra-class triplet embedding over the course of feature learning. With the joint optimization of the improved triplet loss and softmax loss, the multi-task learning is employed to generate more discriminative representation of vehicles. To characterize the intra-class variance, an ideal solution is to adopt exact intrinsic attributes of vehicle images, such as viewpoints, illumination intensity, backgrounds and captured cameras ID. However, it is difficult to explicitly recover these attributes. Alternatively, we resort to clustering to derive group labels, and in particular online clustering method is employed. Moreover, we propose a mean-valued triplet loss [32] to further enhance the learning of discriminative features. Instead of randomly sampling the anchor points, we estimate the positive center of positive samples, such that the impact of improper anchor selection can be eliminated.

### A. Injecting Intra-Class Variance Into the Triplet Loss

*1) Intra-Class Variance Loss:* High inter-class similarity or intra-class variance render learnt features less discriminative. Hence, we propose to inject intra-class variance into triplet embedding to optimize the feature distances between inter-class and intra-class samples via deep metric learning. The design of the loss function is critical. In this work, we employ triplet based deep learning to fulfill metric learning. Specifically, the input is a batch of triplet units $\{< a^p, x^p, x^n >\}$, where $a^p$ is an anchor sample, $x^p$ is a sample belonging to the same vehicle ID with $a^p$, and $x^n$ belongs to the other vehicle ID. The triplet network is to project samples into a feature space where those sample pairs belonging to the same vehicle ID are supposed to be located closer than those from different ones.

To enforce the preservation of relative distances associated with the intrinsic attributes of the instances of each vehicle ID, we incorporate the intra-class variance into the triplet loss (i.e., ICV triplet loss). Specifically, let $a^p$ denote an anchor sample in vehicle $p$'s sample set $S^p$ and $a^{p,g}$ the anchor sample of a group anchor $g$ derived from the vehicle $p$'s sample set $S^{p,g}$. For each vehicle ID, there are one class anchor sample $a^p$ and $G$ group anchors $a^{p,g}$, as illustrated in Fig. 3(b).

For the inter-class relationship, $x^p_* \in S^p$ are positive samples (belonging to the vehicle $p$), and $x^n_* \notin S^p$ are negative samples (not in the vehicle $p$). In terms of intra-class variance, $x^p_g$ and $x^p_i$ denote samples from different groups in vehicle $p$. Then, the inter-class constraint can be formulated as

$$\|f(a^p) - f(x^p_*)\|^2 + \alpha_1 \leq \|f(a^p) - f(x^n_*)\|^2, \tag{2}$$

where $\alpha_1$ is the minimum margin between the samples from different vehicles, $f(x)$ denotes the deep network's feature representation of image $x$. To incorporate the intra-class variance into triplet embedding, the intra-class constraint is further imposed as follows,

$$\|f(a^{p,g}) - f(x^p_g)\|^2 + \alpha_2 \leq \|f(a^{p,g}) - f(x^p_i)\|^2, \tag{3}$$

where $\alpha_2$ is the minimum margin between the samples from different groups within the same vehicle, $x^p_g \in S^{p,g}$ and $x^p_i \notin S^{p,g}$. Amongst the instances with a similar attribute of the same vehicle ID, we set a stronger constraint. Accordingly, we formulate the ICV triplet loss as follows:

$$\begin{aligned} L_{\text{ICV\_Triplet}} &= L_{\text{inter}}(a^p, x^p_*, x^n_*) + \sum_{g=1}^{G} L_{\text{intra}}(a^{p,g}, x^p_g, x^p_i) \\ &= \sum_{x^p_* \in S^p} \frac{1}{2} \max\{\|f(a^p) - f(x^p_*)\|^2 + \alpha_1 - \|f(a^p) \\ &\quad - f(x^n_*)\|^2, 0\} + \sum_{g=1}^{G} \sum_{x^p_g \in S^{p,g}} \frac{1}{2} \max\{\|f(a^{p,g}) - f(x^p_g)\|^2 \\ &\quad + \alpha_2 - \|f(a^{p,g}) - f(x^p_i)\|^2, 0\}, \end{aligned} \tag{4}$$

where $N^p$ and $N^{p,g}$ are the total number of samples in $S^p$ and $S^{p,g}$, respectively. The joint supervision of both intra-class loss ($L_{\text{inter}}$) and inter-class loss ($L_{\text{intra}}$) builds up the group sensitive structure. As such the inter-class constraint as well as intra-class constraint are both incorporated, and the relationship among multiple groups is simultaneously characterized. As illustrated in Fig. 3(b), compared to the original distributions of intra-class samples, with the import of group-wise intra-class constraint, the intra-class samples with similar attributes tend to become more coherent and compact.

*2) Mean-valued Triplet Loss:* The loss function in (4) is sensitive to the selection of anchor $a^p$, and thus choosing improper anchor has a significant influence on the network training. Therefore, instead of randomly selecting anchors from positives in triplet units, we propose the mean-valued triplet loss, to mitigate the impact of the improper anchor selection. Given a positive set $S^p = \{x^p_1, \cdots, x^p_{N^p}\}$ containing $N^p$ positive samples of vehicle $p$, the mean-valued anchor $c^p$ can be formulated as

$$c^p = \frac{1}{N^p} \sum_{n=1}^{N^p} f(x^p_n). \tag{5}$$

Then the mean-valued triplet loss function is defined:

$$\begin{aligned} &L(c^p, S^p, S^n) \\ &= \sum_{k=1}^{N^p} \frac{1}{2} \max\{\|f(x^p_k) - c^p\|^2 + \alpha - \|f(x^n_*) - c^p\|^2, 0\}, \end{aligned} \tag{6}$$

where $x^n_*$ is the negative assigned to the closest anchor $c^p$. If the triplet $< c^p, x^p_k, x^n_* >$ does not satisfy the constraints $\|f(x^p_k) - c^p\|^2 + \alpha \leq \|f(x^n_*) - c^p\|^2$, all the positive samples involving mean value computing are enforced to perform the backward propagation. The partial derivative of positive sample $x^p_k$ with respect to $L(c^p, S^p, S^n)$ is

$$\frac{\partial L}{\partial f(x^p_k)} = f(x^p_k) - c^p + \frac{1}{N^p}(f(x^n_*) - f(x^p_k)). \tag{7}$$

The partial derivative of other positives $x^p_j$ ($j! = k$) is

$$\frac{\partial L}{\partial f(x^p_j)} = \frac{1}{N^p}(f(x^n_*) - f(x^p_k)). \tag{8}$$

The partial derivative of negative samples is:

$$\frac{\partial L}{\partial f(x^n_*)} = c^p - f(x^n_*). \tag{9}$$

It is worth mentioning that our work is related to the center loss [15] and coupled cluster loss [3]. However, the center loss only considers intra-class sample distance, while the coupled cluster loss does not follow that all the positives sampled in computing the center point, should be propagated backwards. By contrast, our mean-valued triplet loss investigates the inter-class distance and intra-class distance simultaneously. To implement the ICV loss, the class anchor $a_p$ and group anchor $a_{p,g}$ in (4) are replaced by class center $c_p$ and the group center $c_{p,g}$, respectively. The class center $c_p$ is the mean value of the total samples of each vehicle ID, and the group center $c_{p,g}$ is the mean value of each group in class $p$. As illustrated in Fig. 3(b), there are one class center and three group centers.

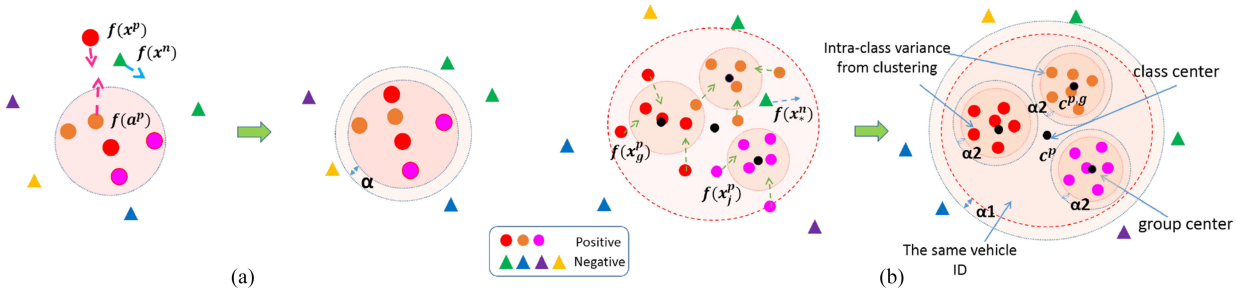



Fig. 3. (a) Illustration of the distance metric optimization process by using the traditional triplet loss and (b) the intra-class variance (ICV) incorporated triplet loss. By grouping the samples by attributes as indicated in different colors in (b), the ICV triplet loss further enforces that the samples of the same group come closer. By contrast, the traditional triplet loss in (a) deals with each category as a whole, and setups the coarse constraints of embedding all the samples into a local space while ignoring the intra-class structure. (better viewed in color).

### B. Online Group Generation

As the feature distribution changes with the network weights updating, we propose to perform online grouping to better characterize the intra-class variance. The group labels are periodically updated online in the training process.

We alternatively update the weights of the network and group labels on the intra-class samples. At the $t$-th iteration, given the vehicle IDs of $I_1, I_2, ...I_N$ and their corresponding sample sets $S^1, S^2, ...S^N$, the group labels are generated. The group label assignment is given by,

$$\mathcal{G}^{(t)} = \{S_{p,g}^{(t)} | g = 1, 2, ...G, \bigcup_{g=1}^{G} S_{p,g}^{(t)} = S^p\},$$

where $S_{p,g}^{(t)}$ is the $g$-th group for vehicle ID $p$ in the $t$-th iteration. When each round of updating groups labels is completed, we fix $\mathcal{G} = \mathcal{G}^{(t)}$ and update the network. Accordingly, the ICV triplet loss function can be further represented as follows:

$$\begin{aligned} L_{\text{ICV\_Triplet}}(f^t) &= L_{\text{inter}}(a^p, x_*^p, x_*^n) + \sum_{g=1}^{G} L_{\text{intra}}(a^{p,g}, x_p^g, x_i^p) \\ &= \sum_{x_*^p \in S^p} \frac{1}{2} \max\{\|f(a^p) - f(x_*^p)\|^2 + \alpha_1 - \|f(a^p) \\ &\quad - f(x_*^n)\|^2, 0\} + \sum_{g=1}^{G} \sum_{x_g^p \in S_{p,g}^{(t)}} \frac{1}{2} \max\{\|f^t(a^{p,g}) \\ &\quad - f^t(x_g^p)\|^2 + \alpha_2 - \|f^t(a^{p,g}) - f^t(x_i^p)\|^2, 0\}, \end{aligned} \tag{10}$$

Then we update group label $\mathcal{G}$ using the k-means clustering

$$\mathcal{G}^{(t+1)} = \arg\min_{\mathcal{G}} \sum_{g=1}^{G} \sum_{x \in S_{p,g}^{(t+1)}} \|f^{(t)}(x) - \mu_g\|^2. \tag{11}$$

where $\mu_g$ is the $g$-th group center. We fix the $t$-th iteration's network parameters, and generate the $t+1$-th iteration's group labels. As that updating labels may cause extra computational cost and frequent updating may lead to slower convergence, we empirically update once every 2 epochs (traverse training data twice). As illustrated in Fig. 5, the vehicles in each resulting group exhibit similar viewpoints (attributes) in the training stage.

### C. Joint Optimization of Multiple Loss

The optimization of the ICV triplet loss alone is inefficient and less effective. First, the ICV triplet loss suffers from the issue of dramatic data expansion. Given a dataset of $N$ images, the number of triplet units is $O(N^3)$, while each iteration takes dozens of triplet units, but only a minority may violate the constraints. As such the convergence for minimizing triplet loss is much slower than other loss constraint (e.g., softmax loss). Second, the triplet loss focuses on similarity distance learning rather than hyperplane decision. Hence, the discriminative power of features are yet to be improved by adding the softmax loss to the loss function. The softmax loss imposes a strong constraint on distinguishing different vehicle IDs. Hence, we employ multi-loss learning to jointly optimize both the ICV triplet loss and softmax loss. By using a hyper parameter $\omega$ to balance two types of loss, the final loss function can be formulated as

$$L_{\text{GSTE}} = \omega L_{\text{softmax}} + (1 - \omega) L_{\text{ICV\_trplet}}. \tag{12}$$

Regarding the hyper parameter, $\omega = 0.75$ works well and is used in our experiments. Fig. 4 illustrates the structure of the deep network with the proposed multi-loss function. In this work VGG_CNN_M_1024 is employed as a base network. It contains 5 convolutional layers and 2 fully-connected layers. The multi-loss works on the last fully-connected layer "$fc7$" with the dimension of 1024. In particular, for the ICV triplet loss, the input feature is $L2$ normalized.

Algorithm 1 shows the optimization pipeline. Given a set of training data, we use mini-batch SGD to optimize the loss function in (6).

## V. Experimental Results

### A. Evaluation Metrics

We adopt two evaluation metrics, mean average precision (mAP) and cumulative match curve (CMC) in our experiments.

*Mean Average Precision:* The mAP metric evaluates the overall performance for re-identification. Average precision is

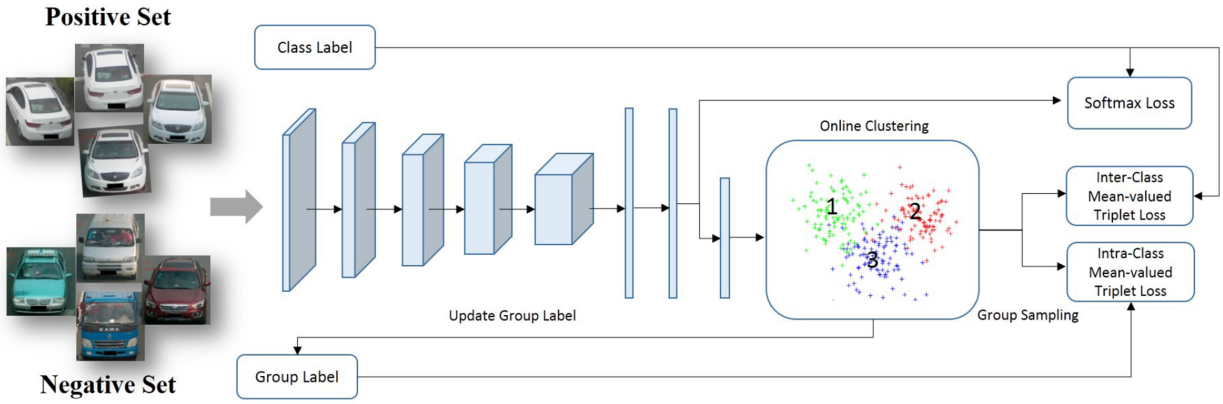



Fig. 4. Illustration of a triplet network by incorporating intra-class variance into triplet embedding, in which the joint learning objective is to minimize the combination of the softmax loss and the triplet loss (consisting of inter-class and intra-class triplet loss).

**Algorithm 1:** Group Sensitive Triplet Embedding

**Input:** Initialized parameters $\theta$ in network layers. Training set $\{S^p|i = 1, 2, ..., N\}$, group number $G$, learning rate $\mu$ group label update interval $m$, training iteration $T$.

**Output:** Learned weights.

```
Group label initial assignment by K-means
for t = 1 to T do
    Sample a mini-batch of training images
    for all S^p in S do
        Compute class center c_p for S^p
        for g = 1 to k in S^g do
            Compute group center c_{p,g}
        end for
        Compute joint loss by L_GSTE
    end for
    Compute total loss L^t in minibatch
    Compute the backpropagation error ∂L^t/∂f(x^t) for each f(x^t)
    Update the θ by θ^{t+1} = θ^t − μ ∂L^t/∂f(x^t) · ∂f(x^t)/∂θ^t
    if t%m = 0 then
        online cluster and update group labels
    end if
end for
```

calculated for each query image as follows:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{\text{gt}}}, \tag{13}$$

where $k$ is the rank in the sequence of retrieved vehicles, $n$ is the number of retrieved vehicles, $N_{\text{gt}}$ is the number of relevant vehicles. $P(k)$ is the precision at cut-off k in the recall list and $gt(k)$ indicates whether the $k$-th recall image is correct or not. Therefore, the mAP is defined as follows:

$$\text{mAP} = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{14}$$

where $Q$ is the number of total query images. Moreover, Top K match rate is also reported in the experiments.

Fig. 5. Exemplar car images from different groups as listed in different columns, which are obtained by applying online clustering ($K = 5$) to the images of a specific car model in VeRI-776 dataset. Each group is with same or similar viewpoints.

*Cumulative Match Characteristics:* The CMC curve shows the probability that a query identity appears in different-sized candidate lists. The cumulate match characteristics at rank $k$ can be calculated as:

$$\text{CMC@}k = \frac{\sum_{q=1}^{Q} gt(q, k)}{Q}, \tag{15}$$

where $gt(q, k)$ equals 1 when the groundtruth of $q$ image appears before rank $k$. The CMC evaluation is valid only if there is only one groundtruth match for a given query.

### B. Datasets

The scale in existing vehicle re-identification datasets cannot provide a sufficient evaluation towards real-world surveillance applications. For example, in Guangdong Province, China, there are more than 30K cameras deployed on the main road. These cameras capture about 43 million vehicle images per day. By contrast, the available well-annotated training set (e.g., VeRI-776 and VehicleID dataset) is terrifically limited. The re-identification methods developed with these databases may not answer the question on the generalization capability.

To meet the emerging demand on large-scale vehicle re-identification, we construct a dataset, namely PKU-Vehicle, which contains tens of millions of vehicle images captured by real-world surveillance cameras in several cities in China. The PKU-Vehicle dataset contains 10 million vehicle images captured from multiple real video surveillance systems across several cities, which in this work serves as a distractor dataset to test the large-scale retrieval performance. Various locations (e.g., highways, streets, intersections), weather conditions (e.g., sunny, rainy, foggy), illuminations (e.g., daytime and evening), shooting angles (e.g., front, side, rear), different resolutions (e.g., 480 P, 640 P, 720 P, 1080 P, 2 K) and hundreds of vehicle brands are involved in PKU-Vehicle dataset. Fig. 7 presents some typical examples from PKU-Vehicle.

Experiments are carried out over four datasets VehicleID [25], VeRI-776 [24], CompCar [1] and PKU-Vehicle. For fair comparison with existing methods, we follow a standard protocol of train/test split.

- VehicleID dataset consists of 221,763 images of 26,267 vehicles (about 250 vehicle models) captured by different surveillance cameras in a city. There are 110,178 images of 13,134 vehicles for training and 111,585 images of 13,133 vehicles for testing. Exactly following the settings in [25], we use three test subsets of different sizes, i.e., 7,332 images of 800 vehicles in small size, 12,995 images of 1600 vehicles in medium size and 20,038 images of 24,000 vehicles in large size.
- VeRI-776 dataset consists of vehicle images captured in a real-word unconstrained traffic scenario, containing about 50,000 images of 776 vehicles, in which each vehicle is captured by 2–18 cameras in different viewpoints, illuminations, resolutions and occlusions. The vehicles are labeled with Bounding Boxes, types, colors, brands and cross-camera relations.
- CompCar dataset, which is a fine-grained vehicle dataset, is mostly collected from Internet. It contains 136,727 vehicle images of 1687 different vehicle models. We select the Part-I subset for training that contains 16,016 images of 431 vehicle models and the remaining 14,939 images for test. It is worth noting that the vehicle images of CompCar used in our experiment are not cropped, and a vehicle occupies about $50 \sim 70\%$ in an image.
- PKU-Vehicle dataset is collected from different surveillance cameras with 10 millions images. The vehicle objects in images are cropped out, such that each image contains one vehicle. In order to thoroughly evaluate the re-identification methods at different scales, we further split the database into eight subsets, i.e., 10 thousands, 50 thousands, 100 thousands, 500 thousands, 1 million, 2 millions, 5 millions, 10 millions.

### C. Experiment Setup

We select the output of L2 Normalization layer as the feature representation for re-identification and retrieval tasks. For fair comparison, we use the VGG_CNN_M_1024 (*VGGM*) [58] as the base network structure, which was also adopted in [25]. In addition, the performance on three other networks Googlenet [35], VGG16 [34], ResNet50 [59] are also reported. All of these networks are initialized with the models pretrained on Imagenet dataset. Regarding the hyper parameters, we set $\alpha = 0.4$ in triplet, and $\alpha_1 = 0.4$, $\alpha_2 = 0.1$ in ICV. Note that the weight $\omega$ in $L_{\text{GS-TRE}}$ is 0.75. The numbers of intra-class groups in CompCar, VeRI-776 and VehicleID are empirically set to be 5, 5 and 2, respectively. Learning rate starts from 0.001 and is divided by 10 every 15 epoches (one forward and backward pass of all the training examples), and the models are trained for 50 epoches. The size of mini-batch, momentum and weight decay is set to 60, 0.9 and 0.0002, respectively. All of the experiments are based on Caffe [60].

TABLE I
RESULTS OF MATCH RATE IN VEHICLE RE-ID TASK IN VEHICLEID DATASET

| Method | | Small | Medium | Large |
|---|---|---|---|---|
| Triplet Loss VGGM [57] | Top 1 | 0.404 | 0.354 | 0.319 |
| CCL VGGM [25] | | 0.436 | 0.370 | 0.329 |
| Mixed Diff + CCL VGGM [25] | | 0.490 | 0.428 | 0.382 |
| ICV triplet loss VGGM | | 0.472 | 0.446 | 0.406 |
| Triplet + Softmax Loss VGGM [48] | | 0.683 | 0.674 | 0.653 |
| GS-TRE loss W/O mean VGGM | | 0.740 | 0.732 | 0.715 |
| GS-TRE loss W/ mean VGGM | | **0.759** | **0.748** | **0.740** |
| Triplet Loss VGGM [57] | Top 5 | 0.617 | 0.546 | 0.503 |
| CCL VGGM [25] | | 0.642 | 0.571 | 0.533 |
| Mixed Diff + CCL [25] | | 0.735 | 0.668 | 0.616 |
| ICV triplet loss VGGM | | 0.738 | 0.713 | 0.665 |
| Triplet + Softmax Loss VGGM [48] | | 0.771 | 0.765 | 0.751 |
| GS-TRE loss W/O mean VGGM | | 0.828 | 0.817 | 0.799 |
| GS-TRE loss W/ mean VGGM | | **0.842** | **0.836** | **0.827** |

To comprehensively evaluate the performance, we provide the baseline and comparison methods as follows: (1) triplet loss [16], (2) triplet + softmax loss [48], (3) mixed Diff + CCL [25], (4) HDC + Contrastive [61], (5) FACT + Plate-SNN + STR [24], (6) GS-TRE loss without a mean-valued anchor for each group, i.e., a randomly selected anchor (GS-TRE loss W/O mean), (6) GS-TRE loss with a mean-valued anchor for each group (GS-TRE loss W/mean).

In the following subsections, we first present and analyze the performance on three different datasets. Subsequently, we discuss the impacts of offline and online grouping in feature learning. Finally, the performance with large-scale distactor dataset PKU-Vehicle is investigated.

### D. Performance Comparisons on VehicleID Dataset

*Re-identification:* Table I presents performance comparisons of the vehicle Re-ID task. The results show that the ICV triplet loss performance generally better as the size of dataset expands. Besides, although ICV triplet loss is worse than Mixes Diff + CCL loss in the top 1 match rate on the small dataset, it achieves a better performance on the top 5 match rate, implying better recall capability benifiting from the intra-class model. The proposed method GS-TRE loss with mean-valued anchors achieves +30% improvements over Mixed Diff + CCL in the large test set. Such significant improvements can be attributed to two aspects. First,

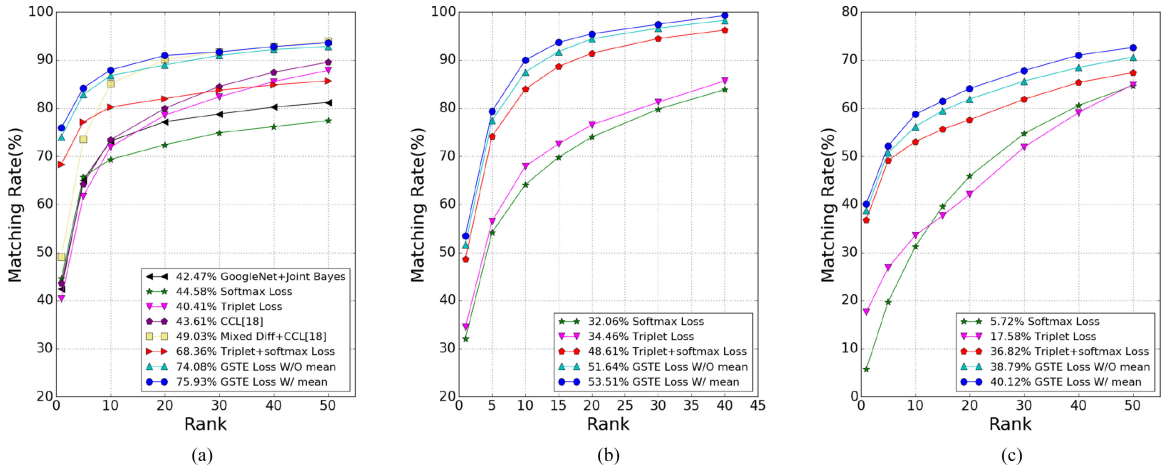



Fig. 6. The CMC curves on (a) VehicleID, (b) VeRI, (c) CompCar datasets. The numbers in the legend of curves are the Top1 value of CMC.


Fig. 7. Samples from the PKU-Vehicle dataset. (a) Vehicles exhibit various shooting angles; (b) Vehicles captured by similar visual appearances; (c) The same vehicle models with different colors; (d) Vehicles with various types; (e) Vehicles captured with occluded and blurred conditions.

we extend the softmax classification to the granularity level of vehicle ID, rather than the vehicle model in [25]. Second, we improve feature learning by introducing the intra-class variance structure and its relevant loss function to triplet embedding. Moreover, as to the Top 1 and Top 5 match rate, our GS-TRE yields significant performance gains compared to the baselines. CMC curves of different methods from Top 1 to 50 on the small test set are given in Fig. 6(a), from which we can observe that GS-TRE shows obvious advantages.

*Retrieval:* Table II lists the retrieval performance comparisons. Note that during the training stage, unlike the methods in [57] and [25] that treat each vehicle model as a category, we treat each vehicle ID as a class (i.e., 13,134 vehicles classes). From Table II, we can observe that simply combining softmax and triplet loss has outperformed Mixed Diff + CCL [25] with significant mAP gain of 19.5% in the large test set. Furthermore, the GS-TRE without mean-valued anchors can further achieve significant improvements across three subsets with different scales. In particular, the mAP improvement on large test set reaches up to 5.8%. Compared to [25], remarkable mAP improvements on large set are observed, i.e., up to 25.3%. It is worth noting that the mean-valued triplet loss in GS-TRE can further obtain about 1.6% mAP gains since the mean values of positives from multiple groups within a vehicle ID yield more reliable anchors, which contributes to better triplet embedding. Fig. 8 shows the feature distribution by t-SNE [62], which demonstrates significantly improved separability brought by GS-TRE learnt feature presentation.

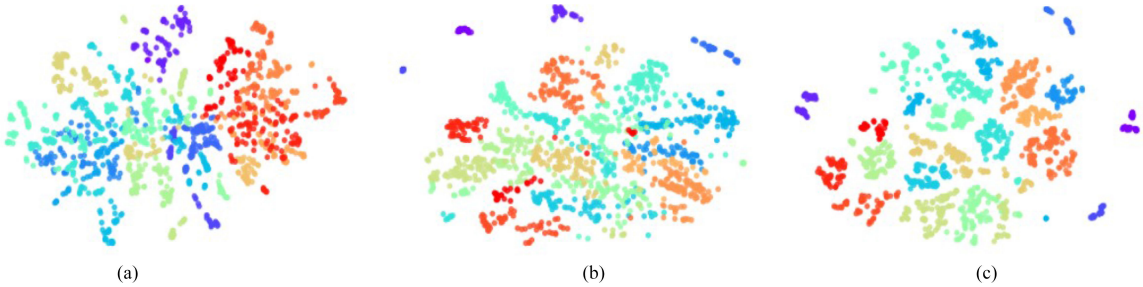
TABLE II
mAP RESULTS OF VEHICLE RETRIEVAL TASK IN VehicleID DATASET

| Methods | Small | Medium | Large |
|---|---|---|---|
| Triplet Loss VGGM [57] | 0.444 | 0.391 | 0.373 |
| CCL VGGM [25] | 0.492 | 0.448 | 0.386 |
| Mixed Diff + CCL VGGM [25] | 0.546 | 0.481 | 0.455 |
| ICV triplet loss VGGM | 0.531 | 0.509 | 0.474 |
| Softmax Loss VGGM | 0.625 | 0.609 | 0.580 |
| HDC + Contrastive [61] | 0.655 | 0.631 | 0.575 |
| Triplet + Softmax Loss VGGM [48] | 0.695 | 0.674 | 0.650 |
| GS-TRE loss W/O mean VGGM | 0.742 | 0.729 | 0.708 |
| GS-TRE loss W/ mean VGGM | **0.754** | **0.743** | **0.724** |

### *E. Performance Comparisons on VeRI-776 Dataset*

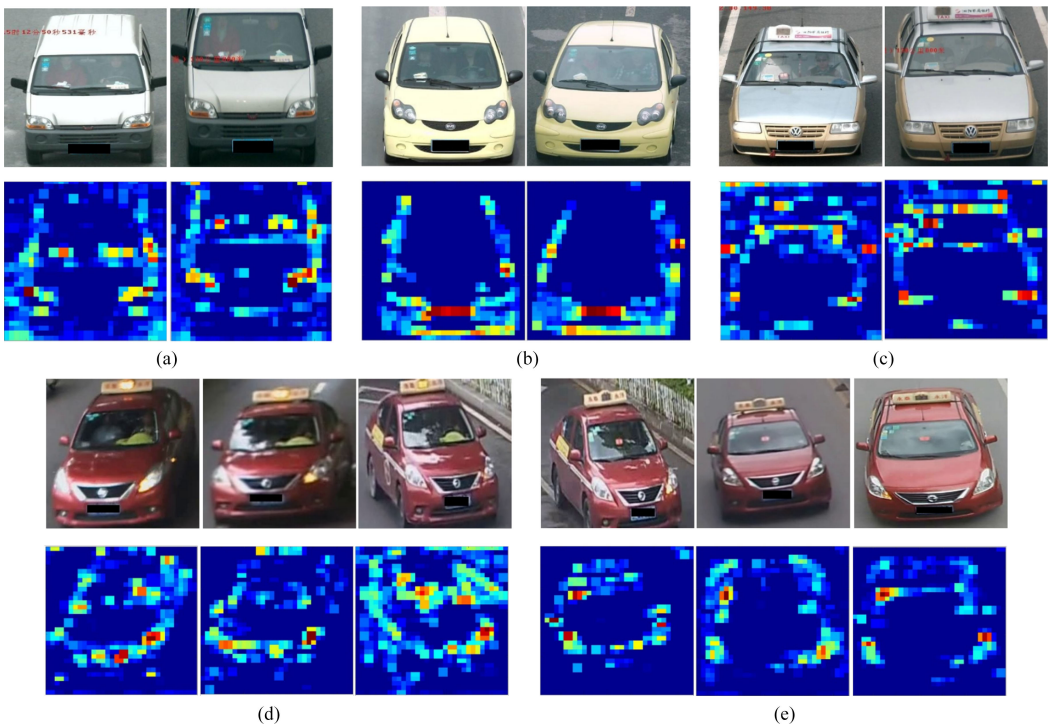
*Retrieval:* We further compare the proposed GS-TRE method with color based feature (BOW-CN), texture feature (LOMO), semantic feature extracted by CNN network (GooleNet, fine-tuned on the CompCars dataset), fusion of attributes and color feature (FACT), Plate recognition trained by SNN model (Plate-SNN), and appearance based coarse filtering (FACT feature), Plate based accurate search (Plate-SNN, Plate-REC), and Spatio-temporal property Based Re-Ranking (STR) mechanism on VeRI-776 dataset.

Table III lists the mAP results on VeRI dataset. The experimental results show that the VGGM network performance by fine training on the VeRI-776 train set (37,781 images of 576 vehicles) significantly outperforms the GoogleNet (much deeper than VGGM) trained with the CompCars dataset (30,955 for



(a) (b) (c)

Fig. 8. Visualization of feature distribution by t-SNE on VehicleID test dataset. Different colors represents different vehicle IDs. We randomly chose 1500 samples from 20 vehicle IDs. The learnt representation by triplet loss can better separate vehicles in feature space than softmax. GS-TRE loss provides an much better feature representation, benefited from the embedding of group structure, and the combination of triplet loss and softmax loss. (a) Softmax, (b) Triplet, (c) GSTE.



(a) (b) (c) (d) (e)

Fig. 9. The visualization of "pool5" feature maps extracted from the VGGM network trained over VehicleID dataset by using the proposed GS-TRE. The vehicle image pairs listed in each subfigure from (a)–(e) are from different vehicle IDs. Noted that there do exist strong response values at the regions containing characteristic details such as headlights, windscreen, decorations, etc. In particular, the annual inspections signs pasted on the top-left corner of the windscreen produce strong responses which helps to distinguish those different IDs of the same vehicle model in practice.

TABLE III
RESULTS OF mAP AND MATCH RATE IN VeRI-776 DATASET

| Methods | mAP | HIT@1 | HIT@5 |
|---|---|---|---|
| BOW-CN [63] | 12.2 | 33.91 | 53.69 |
| LOMO [64] | 9.64 | 25.33 | 46.48 |
| GoogLeNet [1] | 17.04 | 49.82 | 71.16 |
| FACT [4] | 18.49 | 50.95 | 73.48 |
| Plate-SNN [24] | 15.74 | 36.29 | 46.6 |
| FACT + Plate-REC [24] | 18.62 | 51.19 | 73.6 |
| FACT + Plate-SNN [24] | 25.88 | 61.08 | 77.41 |
| FACT + Plate-SNN + STR [24] | 27.77 | 61.44 | 78.78 |
| Softmax Loss VGGM [57] | 34.32 | 83.85 | 92.35 |
| Triplet + softmax loss VGGM [48] | 55.83 | 86.87 | 95.79 |
| GS-TRE loss W/O mean VGGM | 57.76 | 95.79 | 96.45 |
| GS-TRE loss W/ mean VGGM | **59.47** | **96.24** | **98.97** |

the 431 vehicle model). Moreover, we can observe that treating each vehicle ID as the granularity of categorization rather than each vehicle model, can ensure more effective network training, and generate more discriminative feature representation for fine-grained recognition. Fig. 9 visualizes the "pool5" feature maps of examplar vehicles. In Fig. 9 (a)–(c), the extremely similar vehicle pairs belong to different models, but GS-TRE learnt network can generate effective feature maps to distinguish them. In Fig. 9(d) and (e), we present the images of different vehicle ID with the same model, and the feature maps can still produce good responses at characteristic regions (such as windscreen, marks, decorations, etc.), which are effective to distinguish different vehicle.

The FACT feature combines low-level features including color and texture. Liu *et al.* in [24] employs the FACT based


Fig. 10. Exemplar Top 5 retrieval results on VeRi-776 dataset. The images with red box are the wrong results. For each query, the three rows of results from top to down are from the methods of FACT + Plate-SNN + STR [24], Triplet + softmax loss VGGM, and GS-TRE loss W/ mean VGGM.


Fig. 11. Performance curves with the increasing scale of distactor images from PKU-Vehilce. The X axes is in log.

coarse filtering, license plate features based search and spatio-temporal property based re-ranking in vehicle re-identification. In particular, the license plate features in [24] are learnt in deep network with triplet loss. Compared with FACT + Plate-SNN + STR method, we achieve 24.7% mAP improvement which has demonstrated the superiority of the proposed GS-TRE.

Fig. 10 lists Top 5 exemplar retrieval results of FACT + Plate-SNN + STR [24], Triplet + softmax loss VGGM, and GS-TRE loss W/ mean VGGM over VeRI dataset. GS-TRE tends to top rank the recalled images with similar attributes as query images, which is useful to improve the retrieval performance in practice. In view of the comparison results, our method achieve the better mAP and recall performance. In Fig. 12, we provide more results of the Top 10 recall of Triplet + softmax and GS-TRE W/ mean loss VGGM on VeRI dataset. We observe that when the input query is in small resolution (e.g., 180 * 80), the performance would drop due to the difficulties in identifying the characteristics of vehicles. In this scenario, the influence of viewpoint variation on the retrieval performance will also become significant.

*Re-identification:* Fig. 6(b) shows the CMC curve on the VeRI dataset. Note that there is only one groundtruth in reference database as defined in Section VI-A, while in Table III the evaluations of mAP, HIT@1 and HIT@5 are measured with all of the groundtruth of the given query in reference database. From Fig. 6(b) our method achieves consistent improvements over comparison methods (the numbers in legend indicate the CMC value at Top1).

### *F. Performance Comparisons on CompCar Dataset*

Furthermore, we study the effectiveness of our method in CompCar dataset, in which the recognition task is performed

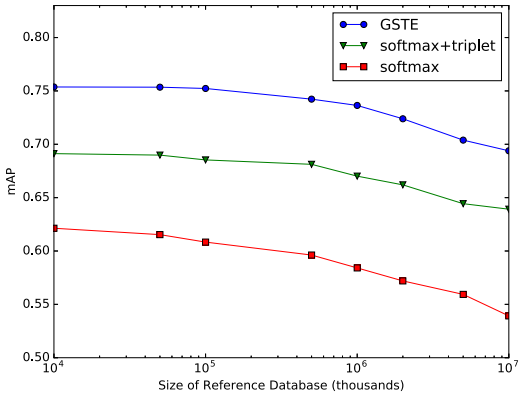



Fig. 12. The visualization of Top 10 retrieval results of a given vehicle. The upper rows are return results by using triplet + softmax VGGM and the bottom rows are our GS-TRE loss W/ mean VGGM method.

TABLE IV
MEAN PRECISION @ $K$ ON COMPCARS RETRIEVAL TASK

| mean precision @ $K$ | 1 | 50 | 500 | All (mAP) |
|---|---|---|---|---|
| Triplet Loss [57] | 0.502 | 0.371 | 0.198 | 0.122 |
| Softmax Loss | 0.456 | 0.282 | 0.167 | 0.091 |
| Triplet + Softmax Loss [48] | 0.719 | 0.586 | 0.419 | 0.349 |
| GS-TRE loss W/O mean | 0.749 | 0.615 | 0.486 | 0.384 |
| GS-TRE loss W/ mean | **0.769** | **0.631** | **0.503** | **0.402** |

at a coarse granularity, i.e., specifying the vehicle model rather than different vehicle IDs.

*Retrieval:* Table IV presents the Top K precision comparisons on CompCar dataset. From the results, the incorporation of intra-class variance into triplet embedding can achieve more than $8.4\%$ precision gains at Top 500 compared with triplet + softmax loss. Overall, the modeling of intra-class variance and its injection into triplet network can significantly improve the discriminative power of feature representation, which plays a significant role in high performance vehicle retrieval.

*Re-identification:* In Fig. 6(c), the triplet loss alone achieves $17.58\%$ match rate at Top-1, and our method brings about $22.5\%$ improvements. Compared with the triplet + softmax method, the proposed method achieves $3.2\%$ higher precision at Top-1 match rate and $5.2\%$ higher at Top-50 match rate, which validates the effectiveness of the GS-TRE.

*Classification:* We also evaluate our method in the classification task. The *VGGM* network is trained with softmax loss with learning rate 0.001 for 40 epoches on ComparCar train set. It yields $78.24\%$ classification accuracy on test set. Further fine-tuning with triplet + softmax loss can bring about $0.7\%$ classification accuracy improvements, while using GS-TRE loss with mean-valued anchors can yield more improvements about $1.7\%$ (i.e., 79.95%). The improvements are less significant compared with re-identification, since the optimization objective mainly works on the feature distance of samples, from which retrieval based tasks can benefit more. Nevertheless, the improvements still demonstrate the effectiveness of preserving intra-class variance that is beneficial in feature learning.

TABLE V
mAP RESULTS OF VEHICLE RETRIEVAL TASK IN VERI DATASET

| VeRI dataset | mAP |
|---|---|
| GS-TRE W/O mean (attribute) | 0.560 |
| GS-TRE W/ mean (attribute) | 0.576 |
| GS-TRE W/O mean (offline) [32] | 0.564 |
| GS-TRE W/ mean (offline) [32] | 0.579 |
| GS-TRE W/O mean (online) | 0.578 |
| GS-TRE W/ mean (online) | **0.594** |

TABLE VI
mAP RESULTS OF VEHICLE RETRIEVAL TASK IN VEHICLEID DATASET

| VehicleID Dataset | Small | Medium | Large |
|---|---|---|---|
| GS-TRE W/O mean (attribute) | 0.735 | 0.723 | 0.702 |
| GS-TRE W/ mean (attribute) | 0.750 | 0.740 | 0.722 |
| GS-TRE W/O mean (offline) [32] | 0.731 | 0.718 | 0.696 |
| GS-TRE W/ mean (offline) [32] | 0.746 | 0.734 | 0.715 |
| GS-TRE W/O mean (online) | 0.742 | 0.729 | 0.708 |
| GS-TRE W/ mean (online) | **0.754** | **0.743** | **0.724** |

### G. Comparisons Over Different Grouping Methods

We thoroughly evaluate the impact of the grouping forms that online versus offline and attributes assignment on the GS-TRE performance. For offline grouping in [32], the images of each vehicle ID are fed into a deep network (VGG_CNN_M_1024) pre-trained on the ImageNet dataset. Then the output of the last fully-connected layer is extracted to perform clustering by K-means. Regarding the attributes assignment, we use the camera IDs in VeRI dataset and viewpoint labels in VehicleID.

Tables V and VI present the comparison results of different grouping methods. The online grouping outperforms the offline method for both with/without mean-valued center methods, since the group labels are periodically updated with the change of feature distributions. Besides, attribute assignment method is better than offline method on VehicleID dataset but worse on VeRi-776 dataset. Since images are captured by 2–18 cameras in VeRI-776 dataset, similar viewpoints for different cameras may exist. Moreover, the performance gain on VeRI dataset are more obvious than VehicleID dataset due to higher intra-class variance on VeRI-776 dataset.

### H. Large Scale Vehicle Retrieval

To extensively investigate the performance in large-scale re-identification task, we conduct experiments with different scales of distractors from the PKU-Vehicle dataset. We select the query and groundtruth from Vehicle ID dataset, which are combined with the distractors from PKU-Vehicle dataset. Eight datasets with the distractor scales of 10 thousands, 50 thousands, 100 thousands, 500 thousands, 1 million, 2 millions, 5 millions, 10 millions are constructed. The mAP performance curves are shown in Fig. 11. The retrieval performance starts to drop from the scale of 100 thousands, and consistently degrades with the increasing scale of distractors. With the 10 million scale of distractors, our method can still achieve 69 % retrieval mAP

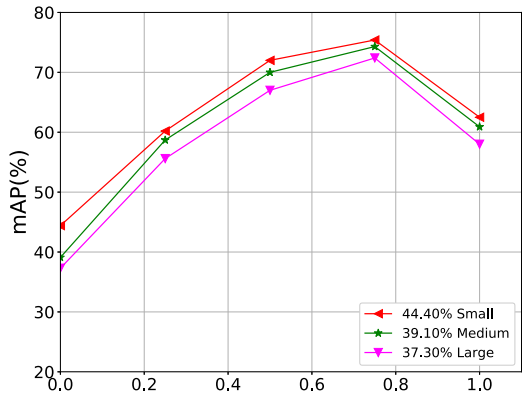



Fig. 13. Performance of GS-TRE W/ mean online on the VehicleID dataset with the variations of $\omega$ in $L_{\text{GS-TRE}}$

TABLE VII
mAP RESULTS OF VEHICLE RETRIEVAL TASK USING DIFFERENT NETWORKS

| Methods | Small | Medium | Large |
|---|---|---|---|
| VGGM Softmax [57] | 0.625 | 0.609 | 0.580 |
| VGG16 Softmax [25] | 0.642 | 0.627 | 0.602 |
| GoogleNet Softmax [61] | 0.647 | 0.624 | 0.603 |
| Resnet Softmax | 0.790 | 0.753 | 0.720 |
| VGGM GS-TRE W/ mean [25] | 0.754 | 0.743 | 0.724 |
| VGG16 GS-TRE W/ mean | 0.768 | 0.752 | 0.733 |
| GoogleNet GS-TRE W/ mean | 0.772 | 0.764 | 0.746 |
| Resnet GS-TRE W/ mean | **0.871** | **0.820** | **0.788** |

with the drop of 6% mAP compared to the original performance with reference database of size 10,000. To the best of our knowledge, this is the first time that the performance evaluation of vehicle re-identification method is carried out over such a large scale dataset, such that the effectiveness of the proposed GS-TRE can be better verified.

### *I. Balance in the Joint Optimization*

The optimization object $L_{\text{GS-TRE}}$ consists of two parts, softmax and ICV triplet loss. When $\omega$ equals to 0, the final loss degenerates as the softmax loss, and on the other extreme when $\omega$ is 1, it turns out to be the ICV triplet loss. More specifically, we provide more results by varying the $\omega$ in Fig. 13. Generally speaking, the combined method is superior to either of these two loss. The hyperparameter $\alpha$ in the margin control also affects the loss convergence. In our experimental results, when $\omega$ is set to be 0.75, the optimal performance can be achieved.

### *J. Comparisons Over Different Networks*

The gains of retrieval performance originate from the proposed GS-TRE loss function, such that GS-TRE should be able to generalize to other network structures. To comprehensively present the superiority of GS-TRE, we extend experiments to the more sophisticated networks. Table VII lists the performance of VGG16, GoogleNet, ResNet50 with GS-TRE loss. Undoubtedly, deeper networks learn better feature representation. From the comparison, the GS-TRE loss based network outperforms the baseline significantly. The improvements across networks suggest that GS-TRE is generic work with the state-of-the-art deep network structure to achieve consistently better performance in vehicle re-identification task.

## VI. CONCLUSION

We present an effective approach to learning discriminative feature representation for vehicle re-identification. In particular, we propose a group sensitive triplet embedding for CNNs to deal with the intra-class variance in learning representation. Moreover, we propose the mean-valued triplet loss to alleviate the negative impact of improper triplet sampling during training stage. Extensive experiments on several benchmarks including VeRI, Vehicle ID, CompCars show that our method can achieve the stage-of-the-art performance. Furthermore, the large-scale vehicle retrieval experiment further demonstrates the effectiveness and robustness of the GS-TRE.

There remain several open issues. Regarding the group generation, it is meaningful to adapt the partition of groups with respect to different iDs, rather than applying a uniform number of clusters. Besides, we may further improve the loss function for vehicle Re-ID, not limited to the global view of vehicle images, which means the discriminative local regions can be located and enhanced feature learning can be done over local regions in a weakly supervised way. It is expected that the combination of the part loss of discriminative regions and the global loss of whole vehicle images may contribute to more effective feature learning.

## REFERENCES

[1] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3973–3981.
[2] J. M. Ernst, J. V. Krogmeier, and D. M. Bullock, "Estimating required probe vehicle re-identification requirements for characterizing link travel times," *IEEE Intell. Trans. Syst. Mag.*, vol. 6, no. 1, pp. 50–58, Spring 2014.
[3] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2167–2175.
[4] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.
[5] D. Zapletal and A. Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2016, pp. 25–31.
[6] Y. Tian, H.-h. Dong, L.-m. Jia, and S.-y. Li, "A vehicle re-identification algorithm based on multi-sensor correlation," *J. Zhejiang Univ. SCI. C*, vol. 15, no. 5, pp. 372–382, 2014.
[7] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* 2016, pp. 1153–1162.
[8] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3716–3724.
[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3498–3505.
[10] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop Fine-Grained Vis. Categorization*, 2011, pp. 1–2.
[11] A. R. Selokar and S. Jain, "Automatic number plate recognition system using a fast stroke-based method," *Int. J. Innovat. Technol. Adapt. Manage.*, vol. 1, no. 7, pp. 34–38, Apr. 2014.
[12] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic license plate recognition (alpr): A state-of-the-art review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 311–325, Feb. 2013.

[13] C. Gou, K. Wang, Y. Yao, and Z. Li, "Vehicle license plate recognition based on extremal regions and restricted boltzmann machines," *IEEE Trans. Intell. Transport. Syst.*, vol. 17, no. 4, pp. 1096–1107, Apr. 2016.

[14] B. Bhattarai, G. Sharma, and F. Jurie, "Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4226–4235.

[15] Y. Wen *et al.*, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[16] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[17] O. Tadmor, Y. Wexler, T. Rosenwein, S. Shalev-Shwartz, and A. Shashua, "Learning a metric embedding for face recognition using the multibatch method," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1396–1397.

[18] E. Simo-Serra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 298–307.

[19] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4004–4012.

[20] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4170–4178.

[21] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3279–3286.

[22] H. Shi *et al.*, "Embedding deep metric for person re-identication a study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 732–748.

[23] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1335–1344.

[24] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 869–884.

[25] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2167–2175.

[26] M. Ye *et al.*, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2553–2566, Dec. 2016.

[27] Y. Wang, J. Choi, V. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1163–1172.

[28] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1114–1123.

[29] M. Cormier, L. W. Sommer, and M. Teutsch, "Low resolution vehicle re-identification based on appearance features for wide area motion imagery," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops.*, 2016, pp. 1–7.

[30] A. Frías-Velázquez, P. Van Hese, A. Pižurica, and W. Philips, "Split-and-match: A bayesian framework for vehicle re-identification in road tunnels," *Eng. Appl. Artif. Intell.*, vol. 45, pp. 220–233, 2015.

[31] Y. Tang, C. Zhang, R. Gu, P. Li, and B. Yang, "Vehicle detection and recognition for intelligent traffic surveillance system.," *Multimedia Tools Appl.*, vol. 76, no. 4, pp. 5817–5832, 2017.

[32] Y. Bai, F. Gao, Y. Lou, S. Wang, T. Huang, and L.-Y. Duan, "Incorporating intra-class variance to fine-grained visual recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 1452–1457.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[35] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[36] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 554–561.

[37] K. Ramnath, S. N. Sinha, R. Szeliski, and E. Hsiao, "Car make and model recognition using 3d curve alignment," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 285–292.

[38] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3d voxel patterns for object category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1903–1911.

[39] R. S. Feris *et al.*, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 28–42, Feb. 2012.

[40] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[41] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep lac: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1666–1674.

[42] K. J. Shih, A. Mallya, S. Singh, and D. Hoiem, "Part localization using multi-proposal consensus for fine-grained categorization," in *Proc. British Mach. Vis. Conf.*, 2015, pp. 128.1–128.12.

[43] C. Huang, Z. He, G. Cao, and W. Cao, "Task-driven progressive part localization for fine-grained object recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2372–2383, Dec. 2016.

[44] X. Liu *et al.*, "Fully convolutional attention networks for fine-grained recognition," in *Proc. 31st AAAI Conf. Artif. Conf.*, 2017, pp. 4190–4196.

[45] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.

[46] J. Krause *et al.*, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 301–320.

[47] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[48] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1386–1393.

[49] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017. [Online]. Available: https://arxiv.org/abs/1703.07737.

[50] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proc. IEEE Int. Comput. Vis. Pattern Recognit.* 2017, pp. 403–412.

[51] X. Yang, M. Wang, R. Hong, Q. Tian, and Y. Rui, "Enhancing person re-identification in a self-trained subspace," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 3, pp. 27:1–27:23, Aug. 2017.

[52] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.

[53] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proc. IEEE Int. Comput. Vision Pattern Recognit.*, 2016, pp. 1114–1123.

[54] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1124–1133.

[55] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 198–207, Jan. 2018.

[56] Z. Li and J. Tang, "Weakly supervised deep metric learning for community-contributed image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1989–1999, Nov. 2015.

[57] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[58] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. British Mach. Vis. Conf.*, 2014, pp. 1–11.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[60] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia.*, 2014, pp. 675–678.

[61] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 814–823.

[62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[63] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[64] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2197–2206.

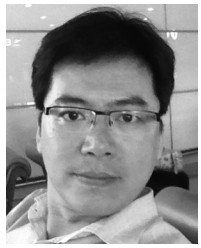
**Yan Bai** (S'18) received the B.S. degree in software engineering from the Dalian University of Technology, Liaoning, China, in 2015. She is currently working toward the M.S. degree at the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. Her research interests include large-scale video retrieval and fine-grained visual recognition.

**Yihang Lou** (S'18) received the B.S. degree in software engineering from the Dalian University of Technology, Liaoning, China, in 2015. He is currently working toward the M.S. degree at the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. His current interests include large-scale video retrieval and object detection.

**Feng Gao** (S'18) received the B.S. degree in computer science from the University College London, London, U.K., in 2007, and the Ph.D. degree in computer science from Peking University, Beijing, China, in 2018. His research interests include the intersection of computer science and art, including but not limited to on artificial intelligence and painting art, deep learning and painting robots, etc.

**Shiqi Wang** (M'15) received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008, and the Ph.D. degree in computer application technology from the Peking University, Beijing, China, in 2014. From 2014 to 2016, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 2016 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong, China. He has proposed more than 30 technical proposals to ISO/MPEG, ITU-T and AVS standards. His research interests include image/video compression, analysis, and quality assessment.

**Yuwei Wu** received the Ph.D. degree in computer science from the Beijing Institute of Technology (BIT), Beijing, China, in 2014. He is currently an Assistant Professor with the School of Computer Science, BIT. From August 2014 to August 2016, he was a Postdoctoral Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and information retrieval. Prof. Wu was a recipient of the outstanding Ph.D. Thesis Award from BIT, and Distinguished Dissertation Award Nominee from the China Association for Artificial Intelligence.



**Ling-Yu Duan** (M'06) received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, NSW, Australia, in 2008. He is a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China, and has served as the Associate Director of the Rapid-Rich Object Search Laboratory, a joint laboratory between Nanyang Technological University (NTU), Singapore, and Peking University (PKU), China, since 2012. Before he joined PKU, he was a Research Scientist with the Institute for Infocomm Research (I2R), Singapore, from March 2003 to August 2008. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics, etc. Prof. Duan was the recipient of the EURASIP Journal on Image and Video Processing Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, China Patent Award for Excellence (2017), and the National Information Technology Standardization Technical Committee "Standardization Work Outstanding Person" Award in 2015. He was a Coeditor of MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13), and is serving as a Co-Chair of MPEG Compact Descriptor for Video Analytics. He is currently an Associate Editor for the *ACM Transactions on Intelligent Systems and Technology* and the *ACM Transactions on Multimedia Computing, Communications, and Applications*. His recent major achievements focus on the topic of compact representation of visual features and high-performance image search.