Attribute and State Guided Structural Embedding Network for Vehicle Re-Identification

Hongchao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo, *Senior Member, IEEE*

***Abstract*— Vehicle re-identification (Re-ID) is a crucial task in smart city and intelligent transportation, aiming to match vehicle images across non-overlapping surveillance camera scenarios. However, the images of different vehicles may have small visual discrepancies when they have the same/similar attributes, *e*.g., the same/similar color, type, and manufacturer. Meanwhile, the images from a vehicle may have large visual discrepancies with different states, *e*.g., different camera views, vehicle viewpoints, and capture time. In this paper, we propose an attribute and state guided structural embedding network (ASSEN) to achieve discriminative feature learning by attribute-based enhancement and state-based weakening for vehicle Re-ID. First, we propose an attribute-based enhancement and expanding module to enhance the discrimination of vehicle features through identity-related attribute information, and we design an attribute-based expanding loss to increase the feature gap between different vehicles. Second, we design a state-based weakening and shrinking module, which not only weakens the state information that interferes with identification but also reduces the intra-class feature gap by a state-based shrinking loss. Third, we propose a global structural embedding module that exploits the attribute information and state information to explore hierarchical relationships between vehicle features, then we use these relationships for feature embedding to learn more robust vehicle features. Extensive experiments on benchmark datasets VeRi-776, VehicleID, and VERI-Wild demonstrate the superior performance and generalization of the proposed method against state-of-the-art vehicle Re-ID methods. The code is available at *https://github.com/ttaalle/fast_assen*.**

***Index Terms*— Vehicle re-identification, attribute-based enhancement, state-based weakening, global structural embedding.**

Manuscript received 5 August 2021; revised 20 June 2022; accepted 19 August 2022. Date of publication 9 September 2022; date of current version 16 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976002, in part by the Natural Science Foundation of Anhui Province under Grant 2208085J18, and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2020-051 and under GXXT-2019-025. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ioannis (Yiannis) Kompatsiaris. *(Corresponding author: Aihua Zheng.)*

Hongchao Li is with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China, and also with the Anhui Provincial Key Laboratory of Network and Information Security, School of Computer and Information, Anhui Normal University, Wuhu 241003, China.

Chenglong Li and Aihua Zheng are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, and the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: ahzheng214@foxmail.com).

Jin Tang and Bin Luo are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China.

Digital Object Identifier 10.1109/TIP.2022.3202370

I. Introduction

VEHICLE Re-identification (Re-ID) aims to identify vehicle images from the gallery images captured from non-overlapping surveillance cameras that share the same identity as the given probe vehicle. It is an active and challenging task and has drawn much attention due to its wide applications in social security, smart city, and intelligent transportation. The blossom of Deep Convolutional Neural Network (DCNN) has witnessed recent breakthroughs in vehicle Re-ID. However, it still faces two severe challenges. 1) The large intra-class discrepancy among the same vehicle images under different states, *e*.g., different camera views, vehicle viewpoints, and capture time as shown in Fig. 1 (a) and (b). 2) The small inter-class discrepancy among different vehicles especially when sharing the same/similar attributes, *e*.g., the same/similar color, type, and manufacturer as shown in Fig. 1 (b), (c) and (d).

Recent efforts have provided various solutions while handling the above challenges. Representative approaches fall into five categories: 1) Global feature based methods [1], [2], [3], [4], [5], [6], which aim to extract the global hand-crafted/deep features of vehicle images by specific metric learning methods. However, global feature based methods are generally hard to capture the intra-class discrepancy and inter-class similarity since only the appearance of vehicle images are considered. 2) Path-based methods [7], [8], [9] usually adopt spatial-temporal information to remove unreasonable vehicles for refining the retrieval results in the inference stage. However, the appearance changes of the vehicle due to spatial-temporal changes are ignored in the learning stage of vehicle features. 3) Viewpoint-based methods [10], [11], [12], which aim to handle viewpoint changes and learn multi-view features via metric learning for vehicle Re-ID. Meanwhile, some viewpoint-based methods [13], [14] generate hard negative cross-view and same-view images for more robust training with a Generative Adversarial Network (GAN) [15]. Although these viewpoint-based methods significantly reduce the intra-class difference, they ignore the intrinsic state factors of vehicles (*e*.g., camera views and capture time) and overlook the challenge of the subtle inter-class discrepancy. 4) Local information enhancement methods [16], [17], [18], [19], [20], [21] usually provide some stable discriminative cues to increase the inter-class discrepancy for vehicle Re-ID. However, local region extraction models usually require a large amount of annotated data which are time and labor consuming. Furthermore, the forthcoming Re-ID model may be sensitive

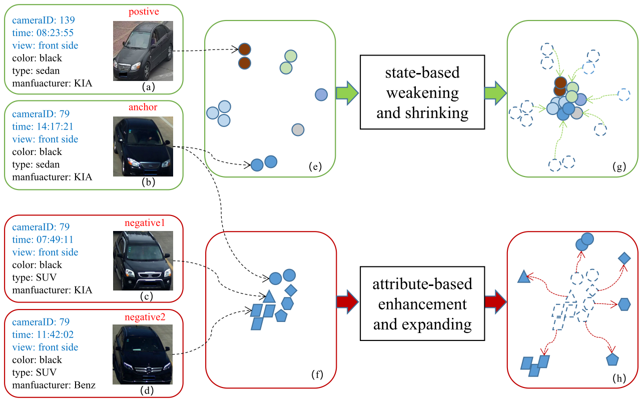


Fig. 1. Illustration of our attribute-based enhancement and state-based weakening framework for vehicle Re-ID. Different colors represent different state information while different markers denote different IDs. In the input image space, vehicle “b, c, d” are more like a same identity vehicle than vehicle “a, b”. Our attribute-based enhancement and expanding module is designed to expand the attribute distribution and re-weight attribute features to the vehicle features to enhance the inter-class difference. For example, the feature distance between manufacture KIA and manufacture Benz is enlarged to force the feature distance between vehicle “b” and vehicle “d” to be greater. In the same way, our state-based weakening and shrinking module is designed to shrink the state distribution and re-weight state features to the vehicle features to weaken the intra-class difference. For example, the feature distance between camera 139 and camera 79 is reduced to force the feature distance between vehicle “b” and vehicle “a” to be smaller. Therefore our attribute-based enhancement and state-based weakening framework can cluster images from the same vehicle compactly and enhance the discrimination between different vehicles.

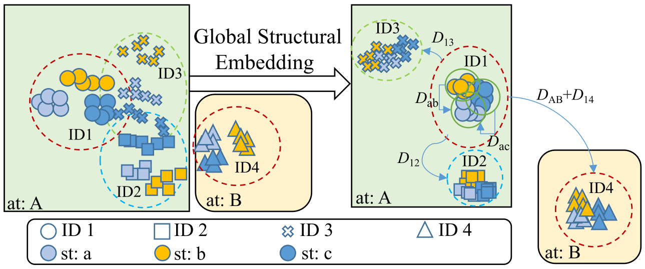


Fig. 2. Illustration of the global structural embedding module for vehicle Re-ID. These points denote the feature embeddings on 60 images from 4 identities in the VeRi-776 testing set. In the input image feature space, vehicle ¡°ID1¡± and ¡°ID3¡± share the same attribute present large overlap due to the large inter-class similarity. Meanwhile, the vehicle ¡°ID1¡± in different states appears sparse feature distribution due to the large intra-class discrepancy. After global structural embedding, images of the same vehicle have been compactly aggregated and the discrimination between different vehicles has been enhanced guided by their state discrepancy ¡°D_{ab}, D_{ac}¡±, instance discrepancy ¡°D_{12}, D_{13}, D_{14}¡±, and attribute discrepancy ¡°D_{AB}¡±.

to the inaccurate part extraction. 5) Attribute-based methods, which use attribute labels to constrain identity features [22], or directly concatenating [23] or summing weighted [24], [25] identity features and attribute features to boost the Re-ID task. Generally speaking, path-based and viewpoint-based methods devote to reduce the impact of identity-unrelated information on vehicle Re-ID, while local information enhancement and attribute-based methods aim to enhance the identity-related information to improve the Re-ID task. In this work, we argue to simultaneously enhance the identity-related and weaken the identity-unrelated information.

In vehicle Re-ID, first, the images of different vehicles with similar attributes share a similar visual appearance (as shown in Fig. 1 (b, c, d)). This results in smaller distances between different vehicles in the feature space (as shown in Fig. 1 (f)), which is the key reason of inter-class similarity in vehicle Re-ID. Therefore, we argue that the feature gap of different vehicle images can be increased by enhancing their identity-related attribute information during feature learning. This is known as knowledge embedding [26] which has been commonly employed in many other computer vision problems [24], [27], [28]. Specifically, we propose an **attribute-based enhancement and expanding module to expand the attribute distribution and re-weight attribute features to the vehicle features to enhance the inter-class difference**. As shown in Fig. 1 (b, d), we enlarge the feature distance between manufacturers “KIA” and “Benz” to force larger feature distance between the two vehicles. Second, the images of the same vehicle (as shown in Fig. 1 (a, b)) under the different states generally present different visual appearance. This results in larger distances between the same vehicle images in the feature space (as shown in Fig. 1 (e)), which is the key reason of intra-class discrepancy in vehicle Re-ID. In the same way, we further argue to decrease the feature gaps between that the images of the same vehicle via state-based weakening during feature learning. Specifically, we propose a **state-based weakening and shrinking module to shrink the state distribution and re-weight state features to the vehicle features to weaken the intra-class difference**. As shown in Fig. 1 (a, b), we reduce the feature distance between camera 139 and camera 79 to encourage the smaller feature distance between the two images. By enforcing the attribute-based enhancement and state-based weakening constraints, identity-related attribute clues will be enhanced while the identity-independent state factors will be weakened in the vehicle features.

Additionally, the deep metric learning methods, which utilize distance metric loss (*e*.g., contrastive loss [29] and triplet loss [5]) rather than cross-entropy loss [29], aim to learn a deep feature embedding space by enforcing the distance between positive pairs smaller than that of negative pairs during learning. However, most exiting metric learning methods only focus on the appearance, which ignores the hierarchical structural relationships caused by the states and attributes. Concretely, different vehicle instances with similar appearance can be further distinguished based on their attribute diversity. Therefore it is effective to consider this relationship to increase the inter-class feature distance as shown in Fig. 2. Meanwhile, the images of the same vehicle instance with large appearance changes can be further recognized by their state information. Therefore, it is useful to decrease the intra-class feature distance between easy and hard positive samples as shown in Fig. 2. Herein, we propose a **global structural embedding module for all vehicle images to cluster images from the same vehicle compactly and enhance the discrimination between different vehicles guided by their state discrepancy, instance discrepancy and attribute discrepancy**.

In this work, we propose an attribute and state guided structural embedding network (ASSEN) towards enlarging the distance of vehicle inter-class features by all available vehicle attributes and reducing the distance of vehicle intra-class features by all available vehicle states. First, we construct an attribute-based enhancement and expanding module to obtain vehicle features enhanced by multiple attributes and design an attribute-based expanding loss to increase the vehicle inter-class gap. Then, we propose a state-based weakening and shrinking module to force the learned vehicle features to weaken state information that interferes with identity and design a state-based shrinking loss to reduce the vehicle intra-class gap. The above two modules encourage our ASSEN to be more focused on its identity-related information rather than identity-unrelated information. Finally, we construct a global structural embedding module to encourage vehicle features to have a global structure related to instance discrepancy, state discrepancy and attribute discrepancy, which can bring hierarchical relationships into the feature embedding to obtain more discriminative vehicle features.

The contributions of this paper can be summarized as follows.

- We design an attribute-based enhancement and expanding module to obtain vehicle features enhanced by multiple attributes. Compare with previous attribute-based Re-ID methods, which use attribute labels to constrain identity features [22], or directly concatenating [23] or summing weighted [24], [25] identity features and attribute features to boost the Re-ID task. Our method utilizes the response relationship between attribute feature and identity feature to highlight the foreground area of the vehicle and expand the subtle differences between the same attribute.
- We propose a state-based weakening and shrinking module to weaken the influence of state information and reduce the state change of the same vehicle. Different from previous work, we further divide the common attribute information into identity-related information and identity-unrelated information. Our key idea is to simultaneously enhance the identity-related and weaken the identity-unrelated information in a unified framework.
- We propose a global structural embedding module to consider hierarchical relationships related to instance discrepancy, state discrepancy and attribute discrepancy in the feature embedding to learn larger weights for hard negative (positive) samples with similar attributes (sharing different states). Existing metric learning methods only consider a small number of samples, or equally treat all samples. Our method adaptively assigns different weights to each sample pair.
- Comprehensive experiments on three large-scale vehicle Re-ID benchmark datasets with or without state and attribute information confirm the effectiveness and generalization of the proposed model.

II. Related Work

We briefly review the related works in the following two folds, *i.e.*, vehicle Re-ID and deep metric learning.

A. Vehicle Re-Identification

Due to wide applications in video surveillance and social security, the vehicle Re-ID task has gained more and more attention in recent years. Liu *et al.* [4] present a deep relative distance learning method to extract both model and instance differences. Features from the model and instance are concatenated to learn the final vehicle feature with vehicle labels. Liu *et al.* [30] fuse color, texture, and deep features for vehicle Re-ID. They show that deep features outperform the others and feature fusion improves the Re-ID performance. Yan *et al.* [31] model the relationship of vehicle images as a multi-grain list to discriminate appearance-similar vehicles. By introducing multi-grain relationships, they force the deep model to learn the more discriminative feature between different grains over many images. Liu *et al.* [7] propose a spatial-temporal relation model to re-rank vehicles to further improve the final results of vehicle Re-ID. Shen *et al.* [8] investigate spatial-temporal association for effectively regularizing vehicle Re-ID results. The spatial-temporal information along the candidate path is effectively incorporated to estimate the validness confidence of the path. Wang *et al.* [32] embed the spatial-temporal regularization into the orientation invariant module for vehicle Re-ID. With spatial-temporal regularization, the log-normal distribution is adopted to model the spatial-temporal constraints and the retrieval results can be refined.

Different from the above global feature based methods and path-based methods, He *et al.* [17] investigate vehicle local regions to learn part-regularized features for vehicle Re-ID. Khorramshahi *et al.* [18] present a dual-path adaptive attention model, to capture key-points related to parts for vehicle Re-ID. Meng *et al.* [19] propose a part perspective transformation on feature space to transform the deformed region to a unified perspective. Liu *et al.* [21] adopt the graph convolutional networks (GCNs) [33] to model the correlation among parts for vehicle Re-ID. However, the part-based approaches need additional part annotations, which takes extra costs. A part prediction network is also needed, which involves more training procedures and complicates the feature extraction model. In addition, identity-related part information is easily disturbed by identity-unrelated information, such as vehicle viewpoints.

To handle the viewpoint variation issue in vehicle Re-ID, Sochor *et al.* [34] learn a 3D orientation vector embedded into the feature map for vehicle recognition. They show that orientation information can decrease classification error and boost verification average precision. Zhou *et al.* [35] generate the opposite side features to handle the viewpoint problem. Zhou *et al.* [13] propose a viewpoint aware network that integrates features from viewpoint-based feature extractors with a GAN to create cross-view features for vehicle Re-ID. Zhou *et al.* [10] exploit the great advantages of DCNN and Long Short-Term Memory (LSTM) [36] to learn transformations across different viewpoints of vehicles. Lou *et al.* [14] propose an embedding adversarial learning network (EALN) to generate hard negative cross-view and same-view images for more robust training in vehicle Re-ID. Jin *et al.* [11] propose an Uncertainty-aware Multi-shot Teacher-Student

(UMTS) Network to exploit the comprehensive information of multi-view of the same vehicle for effective vehicle Re-ID. However, it is difficult to resolve the challenge of vehicle inter-class similarity with these viewpoint learning methods. Most of existing methods only reduce intra-class discrepancy by state (spatial-temporal, viewpoint) information or increase inter-class discrepancy by part information individually, while ignoring the global structural relationship related to states and attributes. We propose an attribute-based enhancement and state-based weakening framework, aiming to explore the global structural relationship to increase the inter-class discrepancy and simultaneously reduce the intra-class discrepancy.

B. Attribute-Based Re-Identification

Recent works in person Re-ID [24], [37], [38], [39] adopt person attributes, such as gender and hair length, as important traits to recognize pedestrians. Khamis *et al.* [37] jointly learn a discriminative projection to a joint appearance-attribute subspace, by effectively leveraging the interaction between attributes and appearance for person Re-ID. Su *et al.* [38] propose a weakly supervised multi-type attribute learning framework based on the triplet loss by pre-training the attributes predictor on independent data. Lin *et al.* [24] simultaneously learn Re-ID embedding and pedestrian attributes, by sharing the same backbone and owning classification FC layers respectively. Sun *et al.* [39] train two different models for attribute and identity recognition tasks and concatenate two branches to one identity vector for Re-ID.

In vehicle Re-ID, Zheng *et al.* [25] propose a deep network architecture guided by meaningful attributes, including vehicle viewpoints, types, and colors, for vehicle Re-ID. Zhao *et al.* [23] collect a new vehicle dataset with 21 classes of structural attributes and proposed a region of interest (ROIs-based) vehicle Re-ID method. Qian *et al.* [22] propose a two-branch stripe-based and attribute-aware deep convolutional neural network (SAN) to learn the efficient feature embedding for vehicle Re-ID task. However, both attributes and vehicle images face challenges caused by appearance changes. Different from previous work, we further divide the common attribute information into identity-related information (named attributes, such as color and type) and identity-unrelated information (named states, such as viewpoint and camera). Our key idea is to simultaneously enhance the identity-related and weaken the identity-unrelated information in a unified framework.

C. Deep Metric Learning

Deep metric learning aims to learn a deep feature embedding space, in which the samples of a same class are close to each other and the samples of different classes are far away. There are two fundamental types of loss functions for deep metric learning, *i*.e., the contrastive loss [29] and the triplet loss [5], which have been widely used in both person and vehicle Re-ID [40], [41], [42], [43]. However, the conventional contrastive loss or triplet loss based deep metric learning often suffers from slow convergence and poor local optima, since only a few samples are considered in each training batch.

There emerge many advances in more robust deep metric learning recently. Chen *et al.* [44] design a quadruplet loss to enforce a larger inter-class variation and a smaller intra-class variation compared to the triplet loss. Sohn *et al.* [45] propose an n-pair loss to generalize triplet loss by allowing joint comparison among more than one negative example. He *et al.* [46] propose a triplet-center loss to learn a center for each class to enhance the discriminative power of the features. Ustinova *et al.* [47] propose a listwise loss to estimate two distributions of similarities between positive (matching) and negative (non-matching) pairs. Wang *et al.* [48] propose a ranked list loss to rank all positive points before the negative points and force a margin between them. Liu *et al.* [49] propose a Group-Group Loss (GGL) to accelerate the intra-group and inter-group feature learning and promote the discriminative ability. Wu propose [50] a margin loss that relaxes unnecessary constraints from traditional contrastive loss and enjoys the flexibility of the triplet loss. However, all the images in positive/negative pairs are treated equally in existing metric learning approaches, which ignore the hierarchical relationships between vehicles. In this paper, we propose a global structural embedding loss to cluster images from the same vehicle compactly and enhance the discrimination between different vehicles guided by their state discrepancy, instance discrepancy and attribute discrepancy.

III. Method

To reduce the intra-class distance of vehicles and increase the inter-class distance of vehicles, we propose an Attribute and State guided Structural Embedding Network (ASSEN). It mainly consists of three modules: attribute-based enhancement and expanding, state-based weakening and shrinking, global structural embedding.

A. Baseline

In this work, our goal is to use the easily obtainable state and attribute information in real-world scenes together with the vehicle ID information to learn the discriminative vehicle identity features. Formally, we denote a vehicle input as $I = \{(x, y^{id}, y_i^{at}|_{i=1}^{M}, y_j^{st}|_{j=1}^{N})\}$, where x and y^{id} denote the input training vehicle image and its associated vehicle identity label. y_i^{at} and y_j^{st} denote the i-th attribute label and the j-th state label of the image x respectively. M and N are the numbers of attribute and state respectively. It's worth noting that, attribute/state labels are not essential during the training since we can use the pre-trained attribute/state branches when the attribute/state labels are absent.

Given a deep backbone network $F(\cdot;\theta)$ with the input image $x \in R^{W\times H\times C}$, where θ represents the learnable parameters of the network. We adopt ResNet-50 [51] without final down-sampling as the backbone model followed by the state-of-the-art vehicle Re-ID methods, such as UMTS [11], PPT [19], FastReID [52], which is also a common setting in person Re-ID methods after PCB [53]. The corresponding vehicle feature tensor encoded by the network is denoted as $T = F(x;\theta) \in R^{w\times h\times c}$. Then the identity classification (cross-entropy) loss $\mathcal{L}_{ce}^{id}$ is in the form of,

$$\mathcal{L}_{ce}^{id} = -y^{id}log(FC(GAP(T))), \tag{1}$$

where GAP denotes a global average pooling operation, and FC denotes a Full Connected layer that predicts the result of classification. In this paper, we regard ResNet-50 with $\mathcal{L}_{ce}^{id}$ as our baseline.

B. Attribute-Based Enhancement and Expanding (AEE) Module

Different from the previous attribute-based Re-ID methods [22], [23], [24], [25], which boost Re-ID tasks by concatenating or weighting attribute features. On the one hand, our AEE module hopes to enhance the image area corresponding to the attribute to improve the feature learning ability of a single sample. On the other hand, our AEE module hopes to expand the distribution of attributes to increase the inter-class distance of samples within a batch.

To obtain the attribute information of the vehicle, we transform the vehicle feature tensor into the vehicle attribute feature tensor. The i-th attribute feature tensor $T_i^{at} \in R^{w\times h\times c}$ can be formulated as:

$$T_i^{at}|_{i=1}^{M} = ReLU(BN(conv_i^{1\times 1}(T))), \tag{2}$$

where $conv_i^{1\times 1}$ denotes 1×1 convolutional operation about the i-th attribute, BN denotes a Bath Normalize operation, and ReLU denotes Rectified Linear Unit. $conv + BN + ReLU$ composes of a common convolutional block in DCNN.

Then the attribute classification loss $\mathcal{L}_{ce}^{st}$ is in the form of,

$$\mathcal{L}_{ce}^{at} = -\sum_{i=1}^{M} y_i^{at}log(FC(GAP(T_i^{at}))), \tag{3}$$

where M is the number of attributes, y_i^{at} denotes the i-th attribute label of the image x.

The attribute tensor will be constrained by the cross-entropy loss and the ground-truth attribute label. Our purpose here is to use attribute labels to enable the output of vehicle features to be guided by multiple attributes. The enhanced tensor can be expressed as:

$$T^e = \frac{1}{M}\sum_{i=1}^{M} T \bigodot Sigmoid(T_i^{at}), \tag{4}$$

where $T^e \in R^{w\times h\times c}$ denotes the attribute enhanced tensor, the $Sigmoid$ function is used to control the value range of T_i^{at} in the interval [0, 1], and $\bigodot$ is the element-wise product. Similar to attention-based Re-ID methods [13], [54], [55], which aims to re-weight the convolutional output of DCNN as a feature combination. However, most of existing attention-based Re-ID methods lack the guidance of identity-related annotations and therefore fail to take advantage of the relationship among the identity, color, and type of the same vehicle. We argue that this intrinsic identity-related information is crucial in vehicle Re-ID.

The overall attribute-based enhancement procedure can be formulated as:

$$T' = T + \beta_1 T^e, \tag{5}$$

where T' denotes the vehicle feature tensor after attribute-based enhancement operation, $\beta_1 = 0.05$ is a hyperparameter used to balance the original feature and the enhanced feature. We add the class activation maps (CAMs) [56] of the attribute (color and type) information, as shown in Fig. 3 (a, b). The color response map and type response map mainly respond to the foreground area related to the vehicle identity, which means that Eq. (5) tends to highlight the foreground area of the vehicle image.

In addition to attribute-based enhancement, we further propose an attribute expanding operation to increase the inter-class attribute discrepancy. The global average pooling (GAP) is used to transfer the i-th attribute tensor $T_i^{at} \in R^{w\times h\times c}$ into the i-th attribute feature vector $f_i^{at} \in R^c$. First, we calculate the i-th attribute standard deviation, which can be formulated as: $D_i^{at} = std(f_i^{at}, \bar{f}_i^{at})$, where $f_i^{at} = GAP(T_i^{at})$ denotes the i-th attribute feature vector about each image in a batch, $\bar{f}_i^{at}$ denotes the i-th attribute mean vector about the whole batch-size. Our purpose here is to expand the feature distribution of the attribute under the premise of attribute classification, thereby increasing the inter-class attribute discrepancy. The attribute-based expanding loss can be formulated as:

$$\mathcal{L}_{ae} = \mathcal{L}_{ce}^{at} + \frac{1}{M}\sum_{i=1}^{M}\frac{1}{1+exp(D_i^{at})}. \tag{6}$$

If there exist two samples that share the same color (or type) in a batch, their color (or type) feature distance will become larger under the premise of classification.

C. State-Based Weakening and Shrinking (SWS) Module

Although attribute-based enhancement and expanding (AEE) module can enhance the inter-class difference by vehicle identity-related attribute information. These identity-related attribute information may be indistinguishable due to diverse state (*e*.g., camera views, vehicle viewpoints, capture time) changes. We argue that merely enhancing identity-related information is not sufficient for Re-ID, weakening the state information that interferes with identification is also crucial for vehicle Re-ID. Herein, we further consider weakening state information to reduce the intra-class feature gap for vehicle Re-ID.

The j-th state feature tensor $T_j^{st} \in R^{w\times h\times c}$ can be formulated as:

$$T_j^{st}|_{i=1}^{N} = ReLU(BN(conv_j^{1\times 1}(T))), \tag{7}$$

where $conv_j^{1\times 1}$ denotes 1×1 convolutional operation about the j-th state. Then the state classification loss $\mathcal{L}_{ce}^{st}$ is in the form of,

$$\mathcal{L}_{ce}^{st} = -\sum_{j=1}^{N} y_j^{st}log(FC(GAP(T_j^{st}))), \tag{8}$$

where N is the number of states, and y_j^{st} denotes the j-th state label of the image x.

The state tensor will be constrained by the cross-entropy loss and the ground-truth state labels. Our goal is to make the

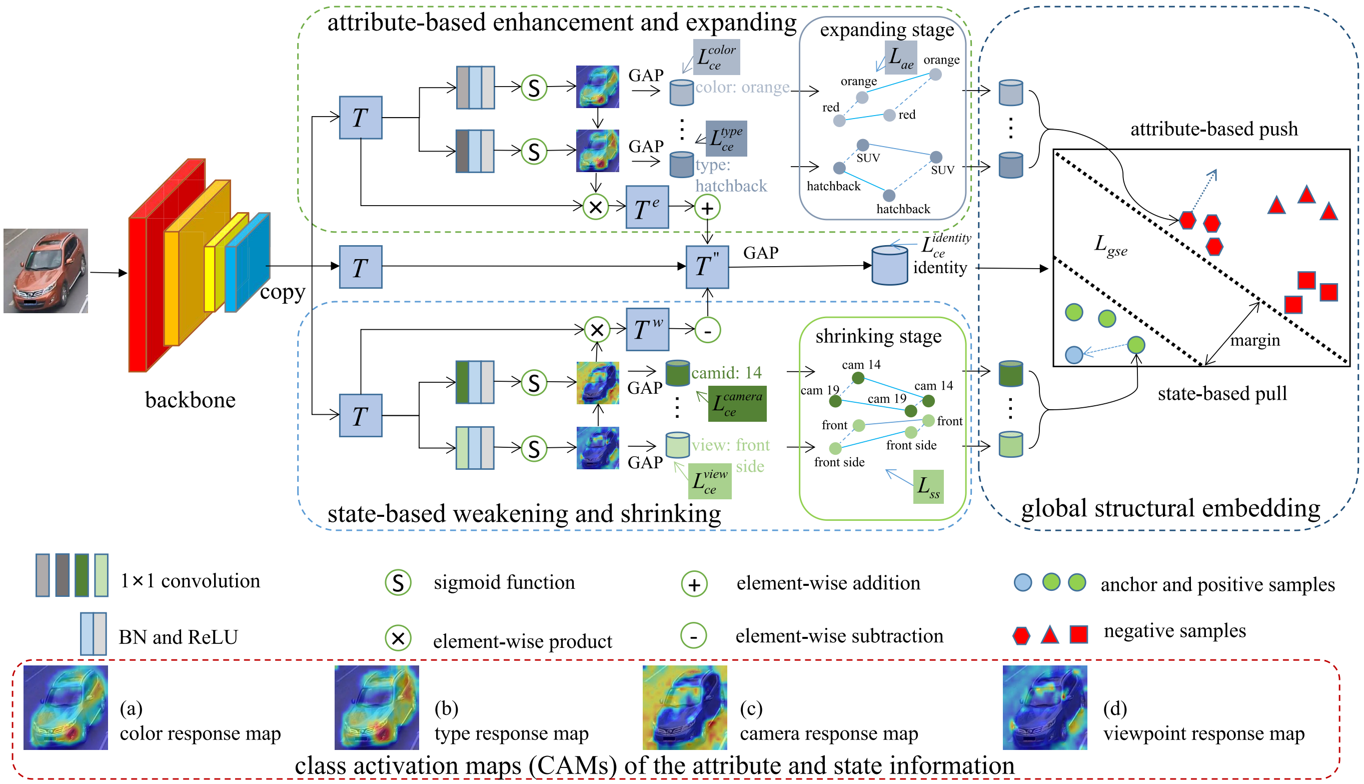


Fig. 3. Pipeline of Attribute and State guided Structural Embedding Network (ASSEN). Given the image x, we first extract the corresponding vehicle feature tensor T via the backbone. Next, we transform the feature tensor T into the attribute-based enhancement and expanding (AEE) module to obtain the enhanced feature tensor T^e. The AEE module is constrained by the attribute-related cross-entropy loss $\mathcal{L}_{ce}^{at}$ and attribute-based expanding loss $\mathcal{L}_{ae}$. Then, we transform the feature tensor T into the state-based weakening and shrinking (SWS) module to obtain the weakened feature tensor T^w. The SWS module is constrained by the state-related cross-entropy loss $\mathcal{L}_{ce}^{st}$ and state-based shrinking loss $\mathcal{L}_{ss}$. Followed by the combination T'' of the feature tensor T, the enhanced feature tensor T^e and the weakened feature tensor T^w to increase the identity-related information and simultaneously reduce the information that interferes with identity. Finally, the global structural embedding (GSE) module embeds instance discrepancy, attribute discrepancy and state discrepancy to obtain more discriminative vehicle features by a hierarchical structure. Note that ASSEN does not require attribute/state labels during the test. Furthermore, attribute/state labels are not essential during the training since we can use the pre-trained attribute/state branches when the attribute/state labels are absent.

learned vehicle feature tensor T alleviate the interference of multiple states as much as possible. The weakened tensor can be expressed as:

$$T^w = \frac{1}{N}\sum_{i=1}^{N} T \odot Sigmoid(T_j^{st}), \tag{9}$$

where $T^w \in R^{w\times h\times c}$ denotes the state weakened tensor, the $Sigmoid$ function is used to control the value range of T_j^{st} to [0, 1], and $\odot$ is the element-wise product.

The overall state-based weakening procedure can be formulated as:

$$T'' = T' - \beta_2 T^w, \tag{10}$$

where T'' denotes the vehicle feature tensor after state-based weakening operation, T' denotes the vehicle feature tensor after attribute-based enhancement operation, $\beta_2 = 0.05$ is a hyperparameter used to balance the original feature and the state weakened feature. We add the class activation maps (CAMs) [56] of the state (camera and viewpoint) information as shown in Fig. 3 (c, d). The camera response map and viewpoint response map mainly respond to the background area of the vehicle image. Therefore Eq. (10) can suppress the background area of the vehicle image.

In addition to designing a state-based weakening procedure, we also added a state-based shrinking operation to reduce the intra-class state discrepancy. The global average pooling (GAP) is used to transfer the j-th state tensor $T_j^{st} \in R^{w\times h\times c}$ into the j-th state feature vector $f_j^{st} \in R^c$. First, we calculate the j-th state standard deviation, which can be formulated as: $D_j^{st} = std(f_j^{st}, \bar{f}_j^{st})$, where $f_j^{st} = GAP(F_j^{st})$ denotes the j-th state feature vector about each image in a batch, $\bar{f}_j^{st}$ denotes the j-th state mean vector about the whole batch-size. Our purpose here is to shrink the feature distribution of the state, thereby reducing the intra-class state discrepancy under the premise of state classification. The state-based shrinking loss can be formulated as:

$$\mathcal{L}_{ss} = \mathcal{L}_{ce}^{st} + \frac{1}{N}\sum_{j=1}^{N}\frac{exp(D_j^{st})}{1+exp(D_j^{st})}, \tag{11}$$

If there exists one sample from different cameras (or viewpoints) in a batch, their camera (or viewpoint) feature distance will become smaller under the premise of classification.

D. Global Structural Embedding (GSE) Module

After attribute-based enhancement and state-based weakening operations, we can obtain a final vehicle feature tensor

$T'' \in R^{w \times h \times c}$. Followed by a global average pooling (GAP) on this tensor, the final vehicle feature vector $f \in R^{c}$ can be expressed as $f = GAP(T'')$. The idea of AEE and SWS is to embed attribute and state information respectively in the training stage to help learn more discriminative identity feature f, which is the feature used in the testing stage.

Although the vehicle feature f can be trained through the cross-entropy loss in Eq. (1), the training and testing of vehicle Re-ID include completely different classes. Therefore it is insufficient to solely rely on the cross-entropy loss. Additionally, the metric learning methods utilize distance metric loss (*e*.g., contrastive loss [29] and triplet loss [5])) to learn a deep feature embedding space where the samples of a same class are close to each other and the samples of different classes are far away. Wu *et al.* [50] propose a simple margin loss that relaxes unnecessary constraints from traditional contrastive loss and enjoys the flexibility of the triplet loss. Based on the margin loss [50], we design a new GSE loss to pay more attention to the hard negative and positive samples by their state discrepancy and attribute discrepancy.

Given a batch of vehicle images $x_i|_{i=1}^{B}$, B is batch size, we can get a batch of vehicle feature vectors $f_i|_{i=1}^{B}$. The margin loss [50] aims to push its negative samples farther than an upper boundary u and pull its positive samples closer than a lower boundary l. Thus $u - l$ is the margin between two boundaries. Mathematically,

$$\mathcal{L}_m = y_{ij} max(d_{ij} - l, 0) + (1 - y_{ij}) max(u - d_{ij}, 0), \quad (12)$$

where $y_{ij} = 1$ if $y_i = y_j$, $y_{ij} = 0$ otherwise. $d_{ij} = \|f_i - f_j\|_2$ is the Euclidean distance between two samples.

It can be seen from Eq. (12) that margin loss only considers the instance difference d_{ij} between sample pairs, but ignores the hierarchical relationship between sample pairs. Concretely, different vehicle instances with similar appearance can be further distinguished based on their attribute diversity, we consider this attribute relationship to help the feature embedding of negative sample pairs:

$$\mathcal{L}_m^- = exp(-d_{ij}^{at})(1 - y_{ij}) max(u - d_{ij}, 0), \quad (13)$$

where d_{ij}^{at} denotes the mean Euclidean distance of the attributes between two negative samples in a batch. It worth noticing that the gradient magnitude concerning any negative embedding is different in Eq. (13). Mathematically,

$$\|\frac{\partial \mathcal{L}_m^-}{\partial f_j}\|_2 = exp(-d_{ij}^{at}), if \;\; y_i \neq y_j, \quad (14)$$

which means that our GSE module encourages negative samples with smaller attribute differences to obtain greater gradient magnitude. If a negative sample pair has the same attribute, the $d_{ij}^{at} \approx 0$, then $exp(-d_{ij}^{at}) d_{ij} \approx d_{ij}$, which denotes the feature embedding mainly depends on the instance difference d_{ij}.

In the same way, since the images of the same vehicle instance with large appearance changes can be further recognized by their state information, we consider this relationship to help the feature embedding of positive sample pairs:

$$\mathcal{L}_m^+ = exp(-\frac{1}{\widetilde{d}_{ij}^{st}}) y_{ij} max(d_{ij} - l, 0), \quad (15)$$

where $\widetilde{d}_{ij}^{st} = d_{ij}^{st} + \epsilon$, $\epsilon = 0.000001$ is a small value to avoid zero denominators, d_{ij}^{st} is the mean Euclidean distance of the states between two positive samples in a batch. $exp(-\frac{1}{\widetilde{d}_{ij}^{st}})$ can be considered as a gradient magnitude of positive embedding, which means that our GSE module encourages positive samples with larger state differences to obtain greater gradient magnitude.

The state and attribute guided global structural embedding loss is:

$$\mathcal{L}_{gse} = S_{ij} y_{ij} max(d_{ij} - l, 0) + W_{ij}(1 - y_{ij}) max(u - d_{ij}, 0), \quad (16)$$

where $S_{ij} = exp(-\frac{1}{d_{ij}^{st} + \epsilon})$ and $W_{ij} = exp(-d_{ij}^{at})$ construct a global structure for the whole batch-size vehicle images. If $S_{ij} = W_{ij} = 1$, $\mathcal{L}_{gse}$ is equivalent to margin loss [50]. $S_{ij} \in [0, 1]$ and $W_{ij} \in [0, 1]$ can be regarded as state-related weights and attribute-related weights respectively.

In GSE module, the designed loss can be explained as giving larger weights for hard negatives and positives. Note that the attribute and state features are imposed into the loss function. The corresponding gradients are as following:

$$\begin{aligned} &\|\frac{\partial \mathcal{L}_m^-}{\partial f_j^{at}}\|_2 = exp(-d_{ij}^{at})(u - d_{ij}), if \;\; y_i \neq y_j, \\ &\|\frac{\partial \mathcal{L}_m^+}{\partial f_j^{st}}\|_2 = exp(-1/\widetilde{d}_{ij}^{st})(d_{ij} - l)/(\widetilde{d}_{ij}^{st} * \widetilde{d}_{ij}^{st}), \;\; else, \end{aligned} \quad (17)$$

which means that our GSE module encourages negative samples with smaller instance differences and attribute differences to obtain greater gradient magnitude of the attribute. Even if two negative samples have the same attributes, the gradient still exists as $\|\frac{\partial \mathcal{L}_m^-}{\partial f_j^{at}}\|_2 = (u - d_{ij})$. Homologous, our GSE module encourages positive samples with larger instance differences and state differences to obtain greater gradient magnitude of the state, until the distance between the positive samples is less than the lower boundary.

To reduce hand-tuned hyperparameters, we reconsider the goals of attribute-based expanding and state-based shrinking, and design a new loss function $\mathcal{L}_{aess}$ to replace the original loss function $\mathcal{L}_{ae}$ and $\mathcal{L}_{ss}$. Mathematically,

$$\begin{aligned} \mathcal{L}_{aess} = &\, \alpha(\mathcal{L}_{ce}^{at} + \mathcal{L}_{ce}^{st}) \\ &+ \frac{\frac{1}{N}\sum_{j=1}^{N} exp(D_j^{st})}{\frac{1}{M}\sum_{i=1}^{M} exp(D_i^{at}) + \frac{1}{N}\sum_{j=1}^{N} exp(D_j^{st})}, \end{aligned} \quad (18)$$

where $\alpha = \frac{2}{M+N}$ is an adaptive parameter inversely proportional to the number of annotations. D_i^{at} (D_j^{st}) represents the i-th attribute (j-th state) standard deviation. Under the premise of attribute/state classification, the attribute difference of all samples is enlarged, while the state difference is reduced. The final objective function for our ASSEN model rewrite as:

$$\mathcal{L}_{total} = \mathcal{L}_{ce}^{id} + \mathcal{L}_{aess} + \eta \mathcal{L}_{gse}, \quad (19)$$

where only η is used to balance the classification learning and metric learning.

IV. Experiment

To validate the superiority of the proposed Attribute and State guided Structural Embedding Network (ASSEN) method, it is compared with state-of-the-art vehicle Re-ID approaches on three large-scale databases.

A. Datasets

VeRi-776 dataset [7] consists of 49357 images of 776 distinct vehicles captured in 20 non-overlapping cameras with various orientations and lighting conditions, where 576 identities with 37778 images and 200 identities with 11579 images are assigned as training and testing respectively. Furthermore, 1678 images from 200 identities have been selected as the queries from the testing set. The original VeRi-776 [7] contains the labels of the vehicle IDs, camera IDs, color IDs and type IDs, while Zheng *et al.* [25] have annotated the viewpoint information, including $front$, $front\_side$, $side$, $rear\_side$, and $rear$. We use two kinds of state information (camera, viewpoint) and two kinds of attribute information (color, type) in VeRi-776 dataset [7].

VERI-Wild dataset [6] is a newly released dataset. Different from VeRi-776 [7] captured at day, VERI-Wild [6] are captured at both day and night. The training subset consists of 277797 images of 30671 vehicles. Besides, there are three different scale testing subsets, *i*.e., Test3000 (Small), Test5000 (Medium), and Test10000 (Large). Except for vehicle ID information, VERI-Wild [6] contains various labels of camera, color, type, and manufacturer annotations. Furthermore, we have annotated the time labels according to the acquisition hour of each image. For example, the image captured at 22:15:29 is annotated as 22, and there are 24 time IDs in total. We use two kinds of state information (camera, time) and three kinds of attribute information (color, type, manufacturer) in VERI-Wild dataset [6].

VehicleID dataset [4] is composed of 221567 images from 26328 unique vehicles. Half of the identities, *i.e.*, 13164, serves for training while the other half for testing evaluation. There are 6 testing splits with various gallery sizes as 800, 1600, 2400, 3200, 6000, and 13164. Following the protocol in [14], [18], and [17], we use the first three splits Test800 (Small), Test1600 (Medium) and Test2400 (Large) for testing. This procedure is repeated ten times and the averaged metrics. Note that VehicleID [4] only contains ID information without any attribute or state information. Therefore, we use the attribute and state branch parameters pre-trained on VERI-Wild [6] to obtain state and attribute information for VehicleID [4].

B. Evaluation Metrics

Following the general evaluation protocols in the Re-ID field [1], [53], [57], the Rank-1 identification rate (R-1), Rank-5 identification rate (R-5), and mean average precision (mAP) are used as performance metrics. Rank-score is an estimation of finding the correct match in the Rank-K returned results. The mAP is a comprehensive index that considers both the precision and recall of the results. To evaluate the stability of our model, we train the model in 10 random trials on each dataset and take the average result as our performance. The corresponding standard deviation values are updated in Table I - IV.

TABLE I

Comparison Results of Our Method Against the State-of-the-Art Methods on VeRi-776 Dataset (in %)

	Methods	mAP	Rank-1	Rank-5	Reference
(1)	BOW-CN [1]	12.2	33.9	53.7	ICCV 2015
	LOMO [2]	9.6	25.3	46.5	CVPR 2015
	GoogLeNet [3]	17.9	52.3	72.2	CVPR 2015
	FACT [30]	18.8	52.2	72.9	ICME 2016
	FDA-Net [6]	55.5	84.3	92.4	CVPR 2019
	FastReID [52]	80.4	96.5	98.4	arXiv 2020
(2)	OIFE [32]	48.0	65.9	-	ICCV 2017
	SCPL [8]	58.3	83.5	90.0	ICCV 2017
	NuFACT [9]	48.5	76.9	91.4	TMM 2018
(3)	VAMI [13]	50.1	77.0	90.8	CVPR 2018
	EALN [14]	57.4	84.4	94.1	TIP 2019
	UMTS [11]	75.9	95.8	-	AAAI 2020
(4)	RAM [16]	61.5	88.6	94.0	ICME 2018
	AAVER [18]	61.2	89.0	94.7	ICCV 2019
	PRN [17]	74.3	94.3	**98.9**	CVPR 2019
	PPT [19]	80.6	96.5	98.3	MM 2020
(5)	DF-CVTC [25]	61.1	91.3	95.8	TETCI 2021
	SAN [22]	72.5	93.3	97.1	MST 2020
	ASSEN	**81.3**$_{\pm 0.2}$	**96.9**$_{\pm 0.1}$	98.7$_{\pm 0.1}$	**Ours**
	¡¡Fast_ASSEN	**¡¡81.7**$_{\pm 0.2}$	**¡¡97.3**$_{\pm 0.1}$	98.8$_{\pm 0.1}$	**¡¡Ours**

TABLE II

Comparison Results of Our Method Against the State-of-the-Art Methods on VehicleID Dataset (in %)

	Methods	Small		Medium		Large	
		R-1	R-5	R-1	R-5	R-1	R-5
(1)	BOW-CN [1]	13.1	22.7	12.9	21.1	10.2	17.9
	LOMO [2]	19.7	32.1	19.0	29.5	15.3	25.6
	GoogLeNet [3]	47.9	67.4	43.5	63.5	38.2	59.5
	DRDL [4]	48.9	66.7	46.4	64.4	41.0	60.0
	FACT [30]	49.5	68.0	44.6	64.2	39.9	60.5
	FDA-Net [6]	-	-	59.8	77.1	55.5	74.7
	FastReID [52]	82.3	95.5	80.7	72.7	77.8	90.1
(2)	OIFE [32]	-	-	-	-	67.0	82.9
	NuFACT [9]	48.9	69.5	43.6	65.3	38.6	60.7
(3)	VAMI [13]	63.1	83.3	52.9	75.1	47.3	70.3
	EALN [14]	75.1	88.1	71.8	83.9	71.0	69.3
	UMTS [11]	80.9	-	78.8	-	76.1	-
(4)	RAM [16]	75.2	91.5	72.3	87.0	67.7	84.5
	AAVER [18]	74.7	93.8	68.6	90.0	63.5	85.6
	PRN [17]	78.4	92.3	75.0	88.3	74.2	86.4
	PPT [19]	79.6	92.3	76.0	89.4	74.8	87.0
(5)	DF-CVTC [25]	75.2	88.1	72.2	84.4	70.5	82.1
	ROIVR [23]	76.1	91.2	73.1	87.5	71.2	84.7
	SAN [22]	79.7	94.3	78.4	91.3	75.6	88.3
	ASSEN	**85.2**$_{\pm 0.2}$	**97.7**$_{\pm 0.1}$	**82.7**	**95.7**	**80.9**	**93.9**
	¡¡Fast_ASSEN	**¡¡86.0**$_{\pm 0.3}$	**97.8**$_{\pm 0.1}$	**¡¡84.5**	**¡¡96.0**	**¡¡82.4**	**¡¡94.3**

C. Implementation Details

1) Network Architecture: We adopt ResNet-50 [51] as the backbone model in our experiments. In our implementation, all the input images are resized to $W \times H \times C = 256 \times 256 \times 3$. Follow [53], we remove the last spatial down-sampling operation in ResNet-50 [51]. After the backbone model, the size of the feature tensor is $w \times h \times c = 16 \times 16 \times 2048$. For classifiers, we use a batch normalization layer [58] and a fully connected layer followed by a softmax function. For data augmentation, the images are augmented with random horizontal flipping, padding 10 pixels, random cropping, and

TABLE III

COMPARISON RESULTS ON MAP OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VERI-WILD DATASET (IN %)

	Methods	Small	Medium	Large	Reference
(1)	GoogLeNet [3]	24.3	24.2	21.5	CVPR 2015
	Triplet [5]	15.7	13.3	9.9	CVPR 2015
	Softmax [7]	26.4	22.7	17.6	ECCV 2016
	DRDL [4]	22.5	19.3	14.8	CVPR 2016
	HDC [61]	29.1	24.8	18.3	ICCV 2017
	Unlabled-GAN [62]	29.9	24.7	18.2	ICCV 2017
	GSTE [41]	31.4	26.2	19.5	TMM 2018
	FDA-Net [6]	35.1	29.8	22.8	CVPR 2019
	FastReID [52]	81.9	75.7	66.7	arXiv 2020
(3)	UMTS [11]	72.7	66.1	54.2	AAAI 2020
(4)	AAVER [18]	62.2	53.7	41.7	ICCV 2019
	PPT [19]	74.2	67.5	59.3	MM 2020
	ASSEN	**80.6**$_{\pm 0.2}$	**74.5**$_{\pm 0.1}$	**66.2**$_{\pm 0.1}$	**Ours**
	¡¡Fast_ASSEN	**¡¡84.3**$_{\pm 0.3}$	**¡¡78.7**$_{\pm 0.2}$	**¡¡70.1**$_{\pm 0.1}$	**¡¡Ours**

TABLE IV

COMPARISON RESULTS ON RANK SCORE OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VERI-WILD DATASET (IN %)

	Methods	Small		Medium		Large	
		R-1	R-5	R-1	R-5	R-1	R-5
(1)	GoogLeNet [3]	57.2	75.1	53.2	71.1	44.6	63.6
	Triplet [5]	44.7	63.3	40.3	59.0	33.5	51.4
	Softmax [7]	53.4	75.0	42.2	69.9	37.9	59.9
	DRDL [4]	57.0	75.0	51.9	71.0	44.6	61.0
	HDC [61]	57.1	78.9	49.6	72.3	44.0	64.9
	Unlabled-GAN [62]	58.1	79.6	51.6	74.4	43.6	65.5
	GSTE [41]	60.5	80.1	52.1	74.9	45.4	66.5
	FDA-Net [6]	64.0	82.8	57.8	78.3	49.4	70.5
	FastReID [52]	96.3	99.2	94.5	98.7	91.1	97.6
(3)	UMTS [11]	84.5	-	79.3	-	72.8	-
(4)	AAVER [18]	75.8	92.7	68.2	88.9	58.7	87.6
	PPT [19]	91.9	97.3	89.1	95.5	84.8	93.2
	ASSEN	**94.9**$_{\pm 0.1}$	**98.3**$_{\pm 0.1}$	**91.7**	**96.5**	**88.8**	**94.7**
	¡¡Fast_ASSEN	**¡¡97.1**$_{\pm 0.1}$	**¡¡99.7**$_{\pm 0.1}$	**¡¡95.6**	**¡¡99.2**	**¡¡93.9**	**¡¡98.4**

random erasing [59]. The Adam optimizer [60] is used with a batch size of 64. We further evaluate our method on a stronger baseline FastReID [52]. Note that due to the GPU memory limitations, we implement FastReID [52] with the same batch-size as our method in 16 $ids * 4\ imgs$ for fair comparison. The new architecture is named Fast_ASSEN in the experiments.

2) Hyper Parameters: In Attribute-based Enhancement and Expanding (AEE) module, β_1 is used to balance the original tensor and the enhanced tensor and set as 0.05. In State-based Weakening and Shrinking (SWS) module, β_2 is used to balance the original tensor and the weakened tensor and set as 0.05. In Global Structural Embedding (GSE) module, we empirically fix the upper and lower boundaries in the GSE module to 1 and 0.3, following the commonly used margin loss [50]. In the final objective function, the weight parameter $\eta = 0.3$, These hyperparameters will be discussed in detail in Table VI. We run our experiments on two Tesla P100 GPU with 16 GB RAM. Our model requires about 13.5 GB of RAM and 348 minutes of training time on VeRi-776 dataset [7]. The base learning rate is 3.5×10^{-4} and the learning rate decays to 3.5×10^{-5} and 3.5×10^{-6} at the 40-th epoch and the 70-th epoch respectively. Our model is trained in a total of 120 epochs.

3) Compared Methods: We compare our method with some state-of-the-art methods which mainly fail into four categories.

a) Global feature based methods: E.g., Bag-of-Words + Color Names (BOW-CN) [1], Local Maximal Occurrence (LOMO) [2], GoogLeNet [3], Fusion of Attributes and Color feaTures (FACT) [30], Feature Distance Adversarial Network (FDA-Net) [6], Deep Relative Distance Learning (DRDL) [4], Triplet [5], Softmax [7], Hard-aware Deeply Cascaded embedding (HDC) [61], Unlabled-GAN [62], Group-sensitive Triplet Embedding (GSTE) [41].

b) Path based methods: E.g., Orientation Invariant Feature Embedding (OIFE) [32], Siamese-CNN + Path + LSTM (SCPL) [8], Null space base Fusion of Attribute and Color feaTures (NuFACT) [9].

c) Viewpoint based methods: E.g., Viewpoint-aware Attentive Multi-view Inference (VAMI) [13], Embedding Adversarial Learning (EALN) [14], Uncertainty-aware Multi-shot Teacher-Student Network (UMTS) [11].

d) Local information enhancement methods: E.g., Region-aware deep Model (RAM) [16], Adaptive Attention Model for Vehicle Re-identification (AAVER) [18], Part-regularized Near-duplicate (PRN) [17], Part Perspective Transformation (PPT) [19].

e) Attribute based methods: E.g., Jointly learns Deep Feature representations, Camera Views, vehicle Types and Colors (DF-CVTC) [25], Two-branch Stripe-based and Attribute-aware Network (SAN) [22], Region of Interests-based Vehicle Re-identification (ROIVR) [23].

D. Comparison With State-of-the-Art Methods

1) Evaluation Results on VeRi-776: Table I reports the performance comparison of our method against the state-of-the-art methods on VeRi-776 dataset [7]. From which we can see, the local information enhancement method PPT [19] has higher performance on VeRi-776 [7] compared with the method UMTS [11] based on viewpoint learning. The reason may be because the viewpoint change of VeRi-776 [7] is not too drastic, challenges mainly come from similar vehicles. Compared with the method based on local information enhancement and viewpoint-based methods, our approach significantly beats the state-of-the-art methods as 81.3% and 96.9% on mAP and the Rank-1 respectively. Although the second-best method PPT [19] achieves 80.6% and 96.5% on mAP and Rank-1 respectively. PPT [19] propose a part perspective transform module to map key points related to part regions to a unified viewpoint on feature space. However, keypoint extraction usually requires a large amount of annotated data which is time and labor consuming, and inaccurate results of keypoint would affect the performance of vehicle Re-ID greatly. Our ASSEN significantly surpasses the most competitive attribute-based method SAN [22] by +8.8% and +3.6% in mAP and Rank-1 accuracies respectively. The key reason is SAN [22] only considers the enhancement of attributes while ignoring the state diversity. By jointly considering the enhancement of attributes, the weakening of states and the hierarchical relationships in the vehicle Re-ID network, our ASSEN learns more robust feature representation on VeRi-776 dataset [7] comparing to the state-of-the-art methods. Fast_ASSEN further boosts the performance in both mAP and ranking scores.

TABLE V

Ablation Study on VeRi-776, VERI-Wild and VehicleID (in %)

	VeRi-776		VehicleID						VERI-Wild					
Variant			Small		Medium		Large		Small		Medium		Large	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
(a1) baseline (L_{ce})	74.3	94.8	85.0	78.6	81.3	74.6	79.6	72.0	72.3	89.5	64.3	84.9	53.6	80.7
(a2) + AEE	76.8	95.5	85.9	82.4	82.0	77.6	80.7	74.7	73.5	92.6	66.4	87.1	57.3	81.1
(a3) + SWS	77.3	95.2	86.1	79.5	83.0	75.4	81.4	72.2	75.6	90.3	69.6	85.2	62.9	80.8
(a4) + GSE	78.9	95.9	88.9	83.4	85.2	80.7	82.8	77.4	77.1	93.2	72.7	89.6	63.8	85.1
(a5) + AEE + SWS	78.3	95.6	87.0	82.6	83.8	79.2	81.3	76.6	76.9	93.5	71.8	89.9	63.1	83.3
(a6) **+ AEE + SWS + GSE**	**81.3**	**96.9**	**90.4**	**85.2**	**88.0**	**82.7**	**85.5**	**80.9**	**80.6**	**94.9**	**74.5**	**91.7**	**66.2**	**88.8**
(b1) FastReID ($L_{ce} + L_{tri}$)	80.4	96.5	85.8	82.3	83.6	80.7	82.6	77.8	81.9	96.3	75.7	94.5	66.7	91.1
(b2) + AEE	80.0	96.8	86.0	83.0	84.0	81.6	82.6	78.2	81.9	96.4	75.2	94.9	66.3	92.2
(b3) + SWS	80.5	96.5	86.6	82.4	84.3	80.7	83.0	77.9	82.2	96.4	76.2	94.6	67.3	91.2
(b4) + AEE + SWS	81.2	96.9	88.1	85.2	86.9	82.6	85.5	81.0	83.0	96.4	78.0	94.9	69.1	92.6
(b5) **+ AEE + SWS + GSE**	**81.7**	**97.3**	**90.9**	**86.0**	**89.1**	**84.5**	**87.2**	**82.4**	**84.3**	**97.1**	**78.7**	**95.6**	**70.1**	**93.9**
(c1) baseline ($L_{ce} + L_{tri}$)	76.6	95.7	85.0	80.2	82.9	77.5	79.7	73.8	76.2	91.8	68.0	87.3	57.8	83.5
(c2) + AEE	76.9	96.3	85.7	81.9	84.0	78.3	80.0	76.1	76.9	93.2	68.9	88.5	59.3	86.6
(c3) + SWS	77.8	95.9	87.3	81.6	85.9	78.3	81.3	75.7	77.9	92.8	72.2	87.8	63.1	84.2
(c4) + AEE + SWS	79.8	96.5	88.7	83.4	87.1	81.0	82.8	78.0	79.0	93.9	73.6	91.0	64.8	88.8
(c5) **+ AEE + SWS + GSE**	**81.3**	**97.0**	**90.4**	**85.4**	**88.6**	**83.6**	**85.9**	**81.2**	**81.0**	**95.4**	**75.2**	**91.9**	**66.8**	**90.2**

2) Evaluation Results on VehicleID: Table II shows the comparison results on VehicleID [4] on three different testing sets. The vehicle images in VehicleID [4] only contain two viewpoints, *e*.g., front and rear, which result in drastic viewpoint changes. As reported in Table II, the method UMTS [11] based on viewpoint learning has higher performance than the local information enhancement method PPT [19] on VehicleID [4] compared with VeRi-776 dataset [7]. This implies that it is necessary to consider joint learning from different viewpoints in VehicleID [4]. In addition to the viewpoint factor similar as UMTS [11], our ASSEN also considers the time factor and the camera factor, as well as the attribute information to enhance the discrimination ability. As shown in Table II, the Rank-1 accuracies of our approach improve 4.3%, 3.9% and 4.8% than UMTS. Note that our methods, ASSEN and Fast_ASSEN, without any attribute and state annotation on VehicleID [4], still significantly beats the state-of-the-art attribute-based methods, especially comparing SAN [22] and ROIVR [23] with additional attribute annotations. This further verifies the generality of our method of leveraging the attribute and state information on more general scenarios.

3) Evaluation Results on VERI-Wild: As shown in Table III and Table IV, our ASSEN achieves competitive results on all of the testing subsets on the VERI-Wild dataset [6]. Specifically, the Rank-1 accuracies of our approach achieve 94.9%, 91.7% and 88.8% on Test3000 (small), Test5000 (middle) and Test10000 (large) respectively, which improve 3.0%, 2.6% and 4.0% than the second-best method PPT [19]. Meanwhile, the mAP of our method achieve 80.6%, 74.5% and 66.2% on Test3000 (small), Test5000 (middle) and Test10000 (large) respectively, which improve 6.4%, 7.0% and 6.9% than the second-best method PPT [19]. The data size of VERI-Wild dataset [6] is about 6 times that of VeRi-776 dataset [7]. Although our Re-ID performance is very close to PPT [19] on VeRi-776 [7], our performance on VERI-Wild dataset [6] is much higher than that of PPT [19], which implies the promising performance in potential large-scale applications. Integrating our method into FastReID [52] consistently improves the performance both mAP and ranking scores.

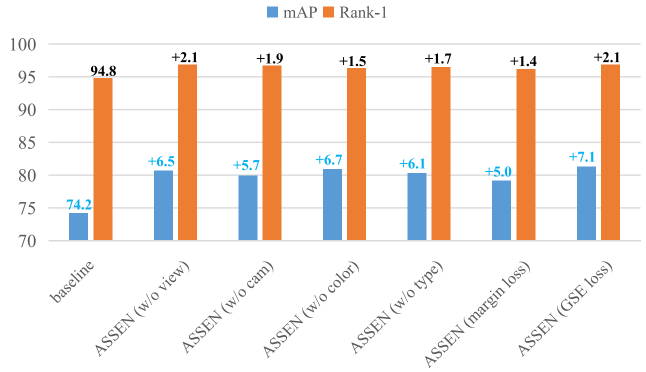


Fig. 4. Subcomponent analysis on VeRi-776.

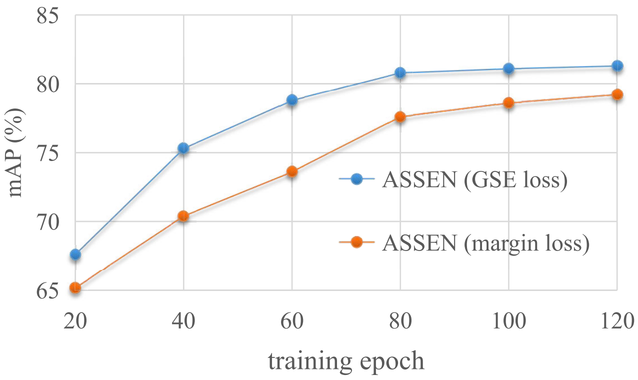
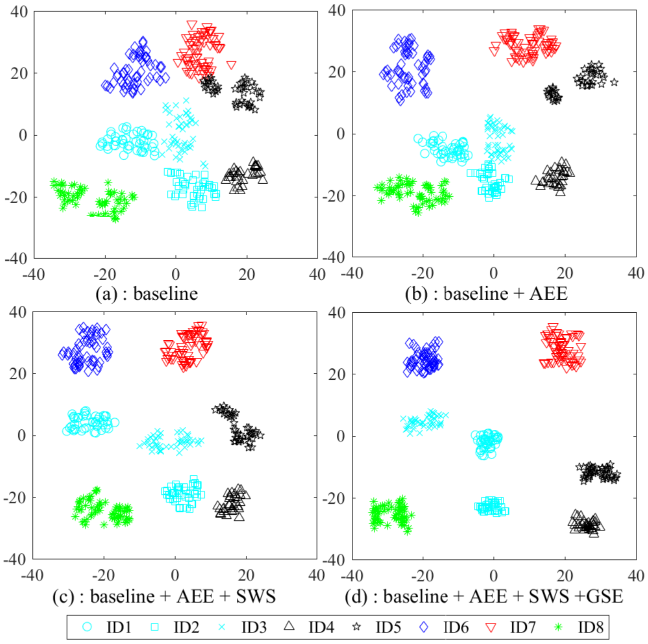


Fig. 5. The mAP performance against the number of training epochs using global structural embedding loss and margin loss [50] on VeRi-776.

E. Ablation Study

1) Component Study: To verify the contribution of the components in our model, we implement several variants of our method on the three datasets, as reported in Table V. Our baseline is ResNet-50 with $\mathcal{L}_{ce}$. By progressively introducing the attribute-based enhancement and expanding module (AEE), state-based weakening and shrinking module (SWS), and global structural embedding module (GSE) into the



(a) : baseline (b) : baseline + AEE (c) : baseline + AEE + SWS (d) : baseline + AEE + SWS +GSE

Fig. 6. T-SNE [63] visualization of the learned feature embeddings on 329 images from 8 identities in the VeRi-776 testing set. The points with the same shape indicate the same identity, while the different colors represent different attributes. These points contain the samples (of ID1, ID2, ID3, ID4) in Fig. 2.

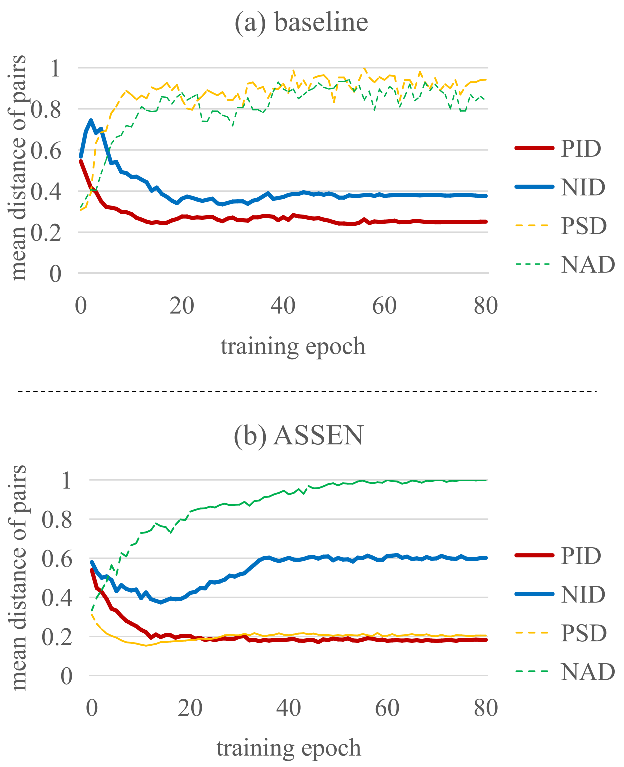


Fig. 7. Feature distance discrepancy of the baseline and ASSEN. Distance discrepancy mainly includes the instance distance between positive sample pairs (PID), the instance distance between negative sample pairs (NID), the state distance between positive sample pairs (PSD) and the attribute distance between negative sample pairs (NAD).

baseline, both mAP, and Rank-1 scores significantly increase on all the three datasets with different test settings. This verifies the contribution of each component in our model.

2) Analysis of Different Baselines: To further validate the effectiveness of our method, we evaluate the component of two stronger baselines, (1) FastReID [52], which is a strong baseline for vehicle Re-ID as shwon in Table V (b1-b5), and (2) the baseline in the state-of-the-art methods such as UMTS [11], PPT [19], FastReID [52], with both cross-entropy loss and triplet loss (baseline (L_{ce} + L_{tri})), as shown in Table V (c1-c5). Note that due to the GPU memory limitations, we implement FastReID [52] with the same batch size as our method in 16 ids∗4 $imgs$ for fair comparison. Consistently, all the AEE, SWS, and GSE modules make effective contributions in our method on the new baselines.

Furthermore, Fig. 6 visualizes the feature map during the ablation study. The AEE module increases the inter-class distance of different attributes, while the SWS module reduces the intra-class distance and increase the inter-class distance with the same attribute. GSE module can further reduce the intra-class gap and increase the inter-class gap.

3) Subcomponent Study: To further evaluate the contribution of each state and attribute, we evaluate our method by removing a certain attribute or state as shown in Fig. 4. It is clear that each attribute or state information contributes to our ASSEN model. In addition, we compare the performance and convergence of the global embedding loss (ASSEN (GSE loss)) with the margin loss [50] (ASSEN (margin loss)) as shown in Fig. 4 and in Fig. 5, respectively. ASSEN (margin loss) denotes $baseline + AEE + SWS + margin\ loss$ and has the same hyperparameters as ASSEN. As shown in Eq. (16), the margin loss [50] can be seen as a special form of our global embedding loss without weight. By considering the hierarchical relationships (inter-class attribute discrepancy and intra-class state discrepancy) between vehicles, our global embedding loss converges faster and achieves better performance.

F. Parameter Analysis

There are five important parameters in our model. β_1 and β_2 balances the contribution of the enhanced feature and the weakened feature respectively, while u and l control the margin between positive samples and negative samples respectively. In the final loss function, η control the weight of classification learning and metric learning. We empirically set $\beta_1 = 0.05$, $\beta_2 = 0.05$, $u = 1$, $l = 0.3$ and $\eta = 0.3$. The parameter analysis results with diverse parameter changes on VeRi-776 [7] are shown in Table VI, which demonstrates that our model is not sensitive to the parameters.

G. Analysis of Distance Discrepancy

To further verify the ability of handling the inter-class similarity and intra-class discrepancy of our method, we visualize the instance distance of positive sample pairs (PID), the instance distance of negative sample pairs (NID), the

TABLE VI
PARAMETER ANALYSIS ON VERI-776 (IN %)

Parameter	Setting	mAP	R-1	Parameter	Setting	mAP	R-1
β_1	0	79.6	95.9	η	0.1	80.8	96.8
	0.05	81.3	96.9		0.2	80.5	97.0
	0.1	80.6	97.0		0.3	81.3	96.9
β_2	0	80.8	96.8		0.4	80.9	96.8
	0.05	81.3	96.9		0.5	80.8	96.9
	0.1	80.1	96.5		0.6	80.6	96.6
u	0.8	79.8	96.1	l	0.2	80.7	96.7
	1.0	81.3	96.9		0.3	81.3	96.9
	1.2	80.3	96.6		0.4	81.0	96.5

state distance of positive sample pairs (PSD) and the attribute distance of negative sample pairs (NAD). We first average the PID/NID/PSD/NAD of each anchor in a batch, and then average over all batches in an epoch. As shown in Fig. 7, our ASSEN significantly shortens the state distance of positive samples (PSD), while increasing the attribute distance of negative samples (NAD), which shortens the instance distance of positive samples (PID) and enlarges the instance distance of negative samples (NID). It shows that weakening the state information can help reduce the intra-class distance, and enhancing the attribute information can help enlarge the inter-class distance. They are both effective ways to improve the discrimination of the vehicle Re-ID network.

V. CONCLUSION

To our best knowledge, this is the first work to solve the problem of Re-ID by enhancing attribute information and weakening state information. In this paper, we first argue the factors that cause the challenge of vehicle Re-ID into state factors and attribute factors. We have contributed an attribute and state guided structural embedding network (ASSEN), followed by three novel modules: attribute-based enhancement and expanding, state-based weakening and shrinking, global structural embedding. Comparing with state-of-the-art vehicle Re-ID methods, extensive experiments demonstrate the promising performance of the proposed method. Although our method requires additional state information and attribute information, this information is easy to obtain and has strong generalization capabilities. In the future, we will consider applying the idea of reducing state discrepancy and increasing attribute discrepancy to other recognition tasks (pedestrians, animals) and unsupervised vehicle Re-ID problems.

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[2] 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. (CVPR)*, Jun. 2015, pp. 2197–2206.

[3] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.

[4] 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. (CVPR)*, Jun. 2016, pp. 2167–2175.

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

[6] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3235–3243.

[7] 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.

[8] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1918–1927.

[9] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.

[10] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3278, Mar. 2018.

[11] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 11165–11172.

[12] A. Porrello, L. Bergamini, and S. Calderara, "Robust re-identification by multiple views knowledge distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 93–110.

[13] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[14] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3794–3807, Aug. 2019.

[15] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[16] X. Liu, S. Zhang, Q. Huang, and W. Gao, "RAM: A region-aware deep model for vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[17] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3997–4005.

[18] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6132–6141.

[19] D. Meng, L. Li, S. Wang, X. Gao, Z.-J. Zha, and Q. Huang, "Fine-grained feature alignment with part perspective transformation for vehicle Reid," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 619–627.

[20] P. Khorramshahi, N. Peri, J.-C. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 369–386.

[21] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 907–915.

[22] J. Qian, W. Jiang, H. Luo, and H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," *Meas. Sci. Technol.*, vol. 31, no. 9, Jun. 2020, Art. no. 095401.

[23] Y. Zhao, C. Shen, H. Wang, and S. Chen, "Structural analysis of attributes for vehicle re-identification and retrieval," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 723–734, Feb. 2020.

[24] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[25] H. Li *et al.*, "Attributes guided feature learning for vehicle re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, early access, Dec. 1, 2021, doi: 10.1109/TETCI.2021.3127906.

[26] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, "Knowledge-embedded representation learning for fine-grained image recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 627–634.

[27] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4190–4196.

[28] L. Lin, L. Huang, T. Chen, Y. Gan, and H. Cheng, "Knowledge-guided recurrent neural network learning for task-oriented action prediction," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 625–630.

[29] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1735–1742.

[30] 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 (ICME)*, Jul. 2016, pp. 1–6.

[31] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 562–570.

[32] Z. Wang *et al.*, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 379–387.

[33] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[34] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3006–3015.

[35] Y. Zhou and L. Shao, "Cross-view GAN based vehicle generation for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 134–146.

[38] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognit.*, vol. 75, pp. 77–89, Mar. 2017.

[39] C. Sun, N. Jiang, L. Zhang, Y. Wang, W. Wu, and Z. Zhou, "Unified framework for joint attribute classification and person re-identification," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 637–647.

[40] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1354–1362.

[41] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[42] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, "Discriminative feature learning with foreground attention for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4671–4684, Dec. 2019.

[43] H.-X. Yu and W.-S. Zheng, "Weakly supervised discriminative feature learning with state information for person identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5527–5537.

[44] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1320–1329.

[45] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857–1865.

[46] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1945–1954.

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

[48] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5207–5216.

[49] X. Liu, S. Zhang, X. Wang, R. Hong, and Q. Tian, "Group-group loss-based global-regional feature learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2638–2652, 2020.

[50] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2840–2848.

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

[52] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A pytorch toolbox for general instance re-identification," 2020, *arXiv:2006.02631*.

[53] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–518.

[54] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[55] H. Guo, K. Zhu, M. Tang, and J. Wang, "Two-level attention network with multi-grain ranking loss for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4328–4338, Sep. 2019.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[57] H. Luo *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[59] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

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

[62] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Hongchao Li received the B.Eng. degree in software engineering and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 2017 and 2022, respectively. He is currently a Lecturer with the School of Computer and Information, Anhui Normal University. His current research interests include person/vehicle re-identification and multimodal learning.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student at the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently an Associate Professor with the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Aihua Zheng received the B.Eng. and master's-doctoral combined program degrees in computer science and technology from Anhui University of China in 2006 and 2008, respectively, and the Ph.D. degree in computer science from the University of Greenwich of U.K. in 2012. She is currently an Associate Professor of artificial intelligence with Anhui University. Her main research interests include computer vision and artificial intelligent, especially on person/vehicle re-identification, audio-visual learning, and multimodal and cross-modal learning.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.

Bin Luo (Senior Member, IEEE) received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University of China in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, U.K., in 2002. He is a Professor with Anhui University, China. He currently chairs the IEEE Hefei Subsection. He has published more than 200 papers in journal and refereed conferences. His current research interests include random graph based pattern recognition, image and graph matching, and spectral analysis. He was a Peer-Reviewer of international academic journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition*, and *Pattern Recognition Letters*.