IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 31, 2019, accepted November 21, 2019, date of publication November 27, 2019,
date of current version December 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956172

A Survey of Vehicle Re-Identification Based on Deep Learning

HONGBO WANG, JIAYING HOU, AND NA CHEN
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
Corresponding author: Hongbo Wang (hbwang@bupt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61002011.

ABSTRACT Vehicle re-identification is one of the core technologies of intelligent transportation systems, and it is crucial for the construction of smart cities. With the rapid development of deep learning, vehicle re-identification technologies have made significant progress in recent years. Therefore, making a comprehensive survey about the vehicle re-identification methods based on deep learning is quite indispensable. There are mainly five types of deep learning-based methods designed for vehicle re-identification, i.e. methods based on local features, methods based on representation learning, methods based on metric learning, methods based on unsupervised learning, and methods based on attention mechanism. The major contributions of our survey come from three aspects. First, we give a comprehensive review of the current five types of deep learning-based methods for vehicle re-identification, and we further compare them from characteristics, advantages, and disadvantages. Second, we sort out vehicle public datasets and compare them from multiple dimensions. Third, we further discuss the challenges and possible research directions of vehicle re-identification in the future based on our survey.

INDEX TERMS Deep learning, intelligent transportation system, vehicle re-identification, vehicle public datasets.

I. INTRODUCTION

In recent years, the development of technology in the field of computer vision and the breakthrough of technology in the field of Internet of Things promote the realization of smart city concept [1]. As important objects in smart city applications, vehicles have attracted extensive attention, a lot of researches about vehicles has been carried out, such as vehicle detection [2], [3], vehicle tracking [4], [5], fine-grained vehicle type recognition [6]–[8], etc. As a frontier and important research topic, vehicle re-identification also caused more and more attention in research area, the purpose of vehicle re-identification is to identify the same vehicle through multiple non-overlapping cameras [9], as shown in Fig. 1.

Vehicle re-identification is one of the core technologies in the intelligent traffic system. Through the ubiquitous monitoring network, a vehicle re-identification system can quickly get the location and time of the target vehicle in the city. With the assistance of the vehicle re-identification system, the target vehicle can be automatically detected, located and tracked across multiple cameras, which saves labor and cost. Besides, vehicle re-identification systems have many possible practical applications, such as intelligent parking, suspicious vehicle tracking, vehicle event detection, vehicle counting, and automatic charging [10]. Furthermore, it has a vital role in applications such as live monitoring or multi-view vehicle tracking for urban surveillance, therefore, vehicle re-identification technology is crucial to the future development of the Internet of things, as well as the construction of intelligent transportation system and smart city.

Although both pedestrians and vehicles are common objects in smart city applications, most attention has been paid to person re-identification in recent years due to the abundance of well-annotated pedestrian data, along with the historical focus of computer vision research on human faces and bodies [120]. Compared to person re-identification, vehicle re-identification is more challenging because of the small inter-class similarity and large intra-class difference. Small inter-class similarity is in reflected images of different cars may look very similar. Vehicles produced by the same or different manufacturers can have similar colors and shapes, so that visual differences between two vehicle images are

The associate editor coordinating the review of this manuscript and approving it for publication was Razi Iqbal.

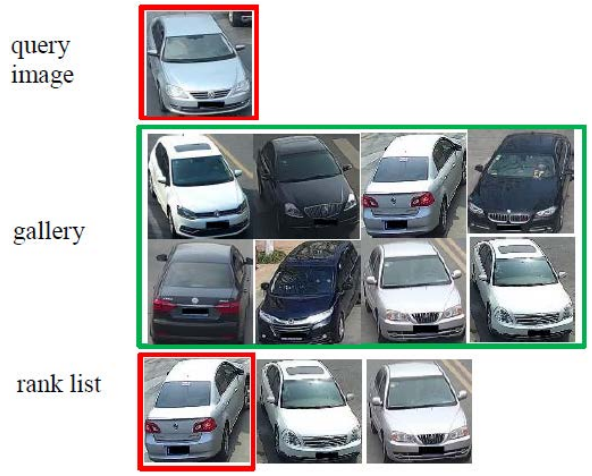

FIGURE 1. Explanation of vehicle re-identification task. The query image is compared with many vehicle images captured by multiple cameras (i.e. gallery), and get a rank list which contains matching vehicle images (images are selected from [92]).

often subtle, making it difficult to distinguish whether the two images belong to the same vehicle. By contrast, people are easier to distinguish because they have more distinct features, including faces and clothes. Large intra-class difference is reflected in images of the same car look different due to the diversity of resolutions, diversity of viewpoints, diversity of illumination and other factors, e.g. visual patterns of vehicle in different viewpoints changes much more than that of people, the images of the same person usually have a common appearance even if there is a large change of viewpoints.

Traditionally, vehicle re-identification problems are solved by combining sensor data with other clues, such as the passing time of a vehicle [11] and wireless magnetic sensors [12]. However, these methods require additional hardware costs and are very sensitive to environmental changes. Besides, because the license plate number is a unique identity of the vehicle, license plate recognition technologies are widely used in vehicle re-identification work [13], [14], that is, identifying license plate numbers of passing vehicles and searching the target vehicle in mass vehicle images by use of license plate numbers. Technologies of license plate recognition is relatively mature at present. But in the real traffic environment, multiple perspectives, lighting, resolution of cameras has obvious influence on the accuracy of license plate recognition, and the license cannot be clearly captured in many cases, e.g. the license plate is blocked, decorated, forged, removed. It is impossible to locate the target vehicle accurately through the retrieval of license plate information. Therefore, vehicle re-identification technologies based on vehicle attributes and appearance characteristics, such as shape, color, texture [15], [16] have received more and more attention. However, these methods have low accuracy, so effectively solve the difficulties and challenges faced by the current research problems and improve the accuracy by using efficient and accurate methods are the research focus in the field of vehicle re-identification. Traditional machine learning adopts hand-crafted features, it is time-consuming and has poor results. With the progress of neural networks in computer vision tasks, methods based on deep learning achieves higher accuracy than previous methods and perform well in real scenes. Therefore, summarize the relevant papers of vehicle re-identification methods based on deep learning is quite necessary and timely.

To the best of our knowledge, there are few comprehensive surveys on the vehicle re-identification are available. Khan *et al.* [17] investigated methods of vehicle re-identification for the first time, which fills the blank of a review paper about vehicle re-identification. They introduced different vehicle re-identification methods including sensor-based methods, hybrid methods, and vision-based methods which are further categorized into hand-crafted feature based methods and deep feature based methods. However, the introduction of vision-based methods is not enough in [17], only 12 papers from 2016 to 2018 are compared. In this paper, vehicle re-identification based on deep learning is divided into five categories, including methods based on local feature, representation learning, metric learning, unsupervised learning, and attention mechanism. Many recent papers are introduced in the introduction of these five categories. Therefore, this paper gives a more comprehensive overview of the methods based on deep learning.

The paper is organized as follows: Section II presents vehicle re-identification methods based on traditional machine learning and deep learning which are further categorized into five categories and gives the comparison between different methods. Section III sorts and compares the current vehicle public datasets and introduces the evaluation strategy of vehicle re-identification, besides, we compare some vehicle re-identification method's accuracy in veri-776 and VehicleID datasets. Section IV discusses the challenges and possible future research directions. Section V, we conclude our work.

II. VEHICLE RE-IDENTIFICATION METHODS

Before the rise of deep learning, traditional machine learning requires hand-crafted features, due to dependence on adjusting parameter manually, only a few parameters are allowed in the design of the feature. After the rise of deep learning, hand-crafted features are no longer needed, instead, it learns features automatically from a lot of training data and contains thousands of parameters so that much time used for designing features manually can be saved and better features can be extracted. Whether hand-crafted features are required is the most obvious difference between deep learning and traditional machine learning. In this chapter, firstly, some methods based on traditional machine learning are introduced, then focusing on methods based on deep learning which include methods based on local features, methods based on representation learning, methods based on metric learning, methods based on unsupervised learning, methods based on attention mechanism and other vehicle re-identification methods.

TABLE 1. Multi-dimensional comparison of vehicle re-identification algorithms based on traditional machine learning.

Method	Characteristics	Advantage	Disadvantage
SIFT	Focus on local key information	1. It is invariant to rotation, scale, and brightness changes 2. Anti-blocking 3. Multi-volume, high speed, and expandable	1. Large amount of calculation 2. Low real-time performance 3. Cannot extract features for smooth edges
HOG	Focus on edge information	1. Ignore the effects of lighting and color on the image 2. Characterization dimensions are reduced	1. Sensitive to occlusion 2. Sensitive to noise
LBP	Focus on texture information	1. Not sensitive to light 2. Simple calculation, fast operation	Sensitive to direction

A. VEHICLE RE-IDENTIFICATION METHODS BASED ON TRADITIONAL MACHINE LEARNING

Traditional machine learning uses feature engineering to artificially refine and clean data. Generally, it includes three steps which are feature extraction, feature coding, and feature classification. There are three methods of feature extraction are widely used, i.e. the scale-invariant feature transform [18] (SIFT), the Histogram of Oriented Gradient [19] (HOG) and the Local Binary Pattern [20] (LBP).

The scale-invariant feature transform (SIFT) feature is a local feature of the images, which maintains the invariance of rotation, scale scaling, and brightness variation. Besides, it maintains a certain degree of stability to the viewing angle change, affine transformation, and noise. SIFT can preserve the uniqueness of the features, and have abundant information. It can be quickly and accurately matched in the massive feature database. In terms of speed, the optimized SIFT matching algorithm has good performance, and can achieve real-time requirements. Besides, SIFT has good scalability and can be conveniently jointed with other forms of feature vectors.

The Histogram of Oriented Gradient (HOG) is a feature descriptor used in the field of computer vision and image processing for target detection. The large-area features are constructed by calculating and counting the gradient direction histograms of the local regions of the image, and overlapping local contrast normalization techniques are used to improve performance. Since the HOG operates on a local grid unit of the image, it can maintain good invariance to both geometric deformation and optical deformation of the image, making both deformations work well for larger spatial fields. That is, the small deformation and optical changes generated in a large area are negligible. Therefore, HOG is particularly suitable for target detection and recognition. At the same time, HOG is not susceptible to noise. Compared with SIFT, HOG is used to describe the entire area, unlike the concept of key points like SIFT. Besides, HOG has no rotation-invariant characteristics. Zapletal and Herout [21] employed the color histogram and the histogram of oriented gradients (HOG) features with linear regression to perform vehicle re-identification. Chen *et al.* [22] proposed a novel grid-based approach to re-identification vehicles grid-by-grid by extracting their HOG features for coarse search and refined the result by using their histograms of matching pairs (HOMs).

The Local Binary Patterns (LBP) is a simple but very efficient texture operator, which compares each pixel to its neighboring pixels and saves the result as a binary number. Its most important attribute is good robustness to grayscale changes caused by factors such as illumination changes. In addition, the calculation of LBP is simple, so that it conducts real-time analysis to an image. Due to its strong discriminating power and simple computational advantages, the LBP is applied in different scenarios in combination with other operators. In [23], Local Variance Measure (VaR) for vehicle re-identification are implemented using Local Binary Patterns (LBP) and joint descriptors. A comparison of the three methods from multi-dimension is shown in Table 1.

In addition to SIFT, HOG, LBP, there are many other well-known operators, e.g. shape context [24], spin image [25], Speeded Up Robust Features [26] (SURF), Space-Time Interest Points [27] (STIP), Histogram of oriented optical flow [28] (HOF), and motion boundary histogram [29] (MBH).

The traditional hand-crafted image features have their characteristics, but their common disadvantage is that the generalization ability is poor, which is manifested in:

1) It is only effective for specific tasks, and cannot be adjusted according to different application scenarios, such as color histogram features. It is effective for image classification tasks, but it does not help the semantic segmentation of images.
2) The features based on hand-crafted only focus on certain aspects of the image, such as the SIFT focuses on the local appearance of the image, the HOG focuses on the image the edge information, the LBP focuses on the texture of the image, etc., and thus the generalization ability is poor.

FIGURE 2. Examples of vehicles with similar appearance but different IDs.

B. VEHICLE RE-IDENTIFICATION METHODS BASED ON DEEP LEARNING

Unlike the traditional machine learning methods described above, deep convolutional neural networks (CNNs) introduce many hidden layers to learn high-level features to improve its generalization ability, not only achieve good performance on the target re-identification task, but also can be well generalized to other computer vision tasks, such as image classification, object detection, semantic segmentation, video tracking, etc. As a result, vehicle re-identification methods based on deep learning become a research hot spot in recent years.

1) VEHICLE RE-IDENTIFICATION METHODS BASED ON LOCAL FEATURES

Due to deep learning and the rapid development of CNNs [30]–[32], significant progress has been made in target re-identification. Because early researches about vehicle re-identification focused on the global feature, that is, using the whole graph to obtain a feature vector for image retrieval. It led to the accuracy bottleneck problem, so some researches begin to pay attention to the local features because differences of similar vehicles are mainly in local areas, as shown in Fig. 2, each column is two similar-looking but different vehicles with different IDs, and the red circle highlights the difference in local areas.

The commonly used methods of extracting local features is to use key point location and region segmentation. The method in [33] that used key point positioning and alignment to extract features of key parts of the object and made detailed comparisons based on key points. Liu *et al.* [34] introduced reinforcement learning to self-adaptive find differentiated regions in fine-grained domains in a weakly supervised manner. Deng *et al.* [35] presented Point Pair Feature Network (PPFNet) for deeply learning a globally informed three-dimensional (3D) local features descriptor which learned local descriptors on pure geometry and was highly aware of the global context.

Methods based on local features have been applied to vehicle re-identification, Wang *et al.* [36] used the method of locating key point and segmenting of different regions to mark the vehicle image as twenty key points and obtained the segmentation results of multiple regions of the target vehicle. They used a convolutional neural network to extract region feature vectors for multiple region segmentation results, and fused with global feature vectors to obtain appearance feature vectors of the target vehicle. Finally, they used the fused feature vector to carry out vehicle re-identification and retrieval, the obtained vehicle appearance features could directly be compared the vehicle appearance features in different vehicle images, and solved the problem that different regions between different vehicle images cannot be compared. Although the scheme considered the influence of attitude on vehicle re-identification, the accuracy of the model was limited by the diversity of the dataset, the dataset needs to include large-scale vehicle images of various angles. In the real world, it is difficult to collect a dataset which includes pictures of the vehicle from different angles as well as the number of pictures reaches hundreds of thousands. In addition, key points need to be labeled for different angles of the vehicle image on the collected dataset, so the number of key points need to be labeled are large which results in a huge workload. Therefore, the method was complicated in terms of feasibility and workload.

More scholars studied methods based on local features. Because it is hard to distinguish vehicles that share the same model and maker only by global feature because they are similar in global appearance, some methods combine local features and global features for vehicle re-identification. Liu *et al.* [37] proposed a Region-Aware deep Model (RAM), which extracted features from local regions instead of only extracting global features, RAM embed the detailed visual cues in local regions as each local region conveyed more distinctive visual cues. Besides, they introduced a new algorithm that jointly used vehicle IDs, types and colors to train the model, which fused more cues for training, resulting in more discriminative global feature sand regional features. There is a similar approach, He *et al.* [38] developed a novel framework which was trained end-to-end with combined local and global constraints by introducing a detection branch. A local module focused on the part features to distinguish the subtle discrepancy in visual features, parts included front and back lights, front and back windows, and vehicle brand. A global module was regularized by the part attentions in the local module. The part-regularized discriminative feature preserving method enhanced the perceptive ability of subtle discrepancies.

Local information is important in vehicle re-identification. To localize local regions that contain more distinctive visual cue, Peng *et al.* [39] proposed a Multi-Region Model (MRM) to extract features from a series of local regions, for each local region, a Spatial Transformer Network (STN) based localization model was introduced. They presented a context-based re-ranking method, the method generated the re-ranking list by combing context and content to measure the similarity between neighbors, and the method improved the accuracy of vehicle re-identification. There are same similar approaches. Chen *et al.* [40] proposed an end-to-end trainable two-branch Partition and Reunion Network (PRN), which combined

global and local features together to build more robust visual signatures. Since salient local information is important in vehicle re-identification, they adopted multiple partitions along three dimensions (height, width, and channel) in feature maps to extract more local features from each dimension of images. But because height and width belong to spatial dimensions, features extracted from the same location on the feature map could be considered twice, so the network was split into one height-channel branch and one width-channel branch to avoid certain spatial features being considered twice. Besides, Zhao *et al.* [41] proposed a region of interests (ROIs)-based vehicle re-identification method, which extracted deep features from the classification model and used the results of the single shot multibox detector (SSD). Local features of ROIs could be extracted according to the detected location, these ROI features that were combined into a structural feature could mark a vehicle uniquely. The uniqueness of this method lies in the combination of a classification model and a detection model to solve the problem of vehicle re-identification. Ma *et al.* [42] proposed a refined part model to learn an efficient feature embedding. The refined part model was formed through a Grid Spatial Transformer Network (GSTN), and it could automatically locate the vehicle and perform division for local features. Besides, residual attention was conducted to give an additional refinement for a fine-grained identification, the refined part features were fused to form an efficient feature embedding finally, so that improved the accuracy of vehicle re-identification.

In summary, the advantages of methods based on local features are reflected in it can capture unique visual clues conveyed by local areas and improve the perception of nuance, which helps a lot to distinguish between different vehicles and improve the accuracy of vehicle re-identification. Besides, many researchers combine local features with global features to improve the accuracy of vehicle re-identification. Hoverer, the disadvantage of methods based on local features is the extraction of local features will significantly increase the computational burden.

2) VEHICLE RE-IDENTIFICATION METHODS BASED ON REPRESENTATION LEARNING

In the real application scenario of vehicle re-identification, the significant changes in camera shooting angle may lead to significant differences in local key areas. It is difficult to achieve high accuracy by vehicle re-identification only by local features. Due to the rapid development of CNNs, significant progress has been made in representation learning (feature learning) [43], the representations are formed by the composition of multiple non-linear transformations of the input data to yield abstract and useful representations for classification, prediction and other tasks [43]. Representation learning aims to get a valid representation of the data by training large amounts of data, making it easier to extract useful information when building classifiers or other predictors. Specifically, using CNNs to train a large amount of data, feature extraction is automatically performed from the image according to different task requirements such as classification and recognition. Representation learning is a very important method in the field of re-identification, it has high robustness and stable training, and has been applied to person re-identification [44]. Therefore, some jobs applied representation learning to the solving of vehicle re-identification.

It is very important to learn more discriminative representations from the vehicle appearance, Zheng *et al.* [45] proposed DF-CVTC, a unified deep convolutional framework to jointly learned deep feature representations guided by the meaningful attributes, including camera views, vehicle types, and colors for vehicle re-identification. These components were collaborative to each other, and thus improved the discrimination ability of the learned representations, besides, VS-GAN, a vehicle generation model was developed to enhance the diversity of view data. A Deep Feature Fusion with Multiple Granularity (DFFMG) method for Vehicle re-identification was proposed in [46], it used both global feature and part feature fusion, partitioned vehicle images along with two directions (i.e. vertically and horizontally) and integrated discriminative information with various granularity. DFFMG consisted of one branch for global feature representations, two for vertical local features representations and other two for horizontal local features representations.

Some methods based on representation learning have novel and unique ideas. Hou *et al.* [47] proposed a random occlusion assisted deep representation learning based vehicle re-identification algorithm. What's unique about this algorithm was that it employed the random occlusion method to randomly occlude the original training images, which simulated some occlusion situations in the real world to a certain degree. Moreover, it increased the number of training samples and prevented the model from over-fitting, then, the joint identification and verification learning optimization were performed on training the original images and occluded images through the developed network. Krause *et al.* [48] thought modeling objects as two-dimensional representations of a collection of unconnected views limited their ability to generalize across viewpoints, so they lifted two state-of-the-art two-dimensional object representations to three-dimensional on the level of appearance and location. Three-dimensional object representations have been widely used in the context of multi-view object class detection and scene understanding, but have not yet widely used in fine-grained categorization, they provided first experimental results on the challenging task of three-dimensional reconstruction of fine-grained categories and showed their 3D object representations outperform their state-of-the-art two-dimensional counterparts for fine-grained categorization. A framework based on deep learning which could lead to an efficient representation of vehicles was proposed in [49], the key of the framework was that learning variational feature was employed to generate variational features which were more discriminating and long short-term memory (LSTM) was used to learn the relationship among different viewpoints of a vehicle. The advantage of the framework was that it can be derived highly

discriminating representations for vehicle images improved the performance of vehicle re-identification. Besides, it is believed that the idea of using variational feature learning with Kullback-Leibler Divergence can not only improve the performance of vehicle re-identification but also can improve the quality of object representation on other similar scenes.

There are other methods based on representation learning. To address issues such as data labeling, visual domain mismatch between datasets and diverse appearance of the same vehicle, Wu *et al.* [50] proposed a CNN-based vehicle re-identification system, the adaptive representation learning technique based on the space-time prior was used to automatically get positive and negative training samples from unlabeled testing videos. They trained a vehicle feature extractor in a multi-task learning manner and fine-tuned the feature extractor on the target domain so that the deep network could adapt to the visual domain of the testing videos. To accelerate the procedure of representation learning, a new distance loss was proposed in [51], it considered samples of the identical vehicle as an image set, and it pulled samples in the same set close to each other and pushed different sets away from each other, using this way to guide the network training procedure to optimize the distance between and within image sets, advantage of the proposed loss lay in better efficiency than the commonly used sample-wise triplet loss. Jiang *et al.* [52] presented a multi-attribute driven vehicle re-identification approach which consisted of a multi-branch architecture and a re-ranking strategy to learn discriminative representations. The multi-branch architecture explicitly leveraged the vehicle attribute cues such as color, model to enhance the generalization ability. The re-ranking strategy introduced the spatial-temporal relationship among vehicles from multiple cameras to construct the similar appearance sets and utilized Jaccard distance between these similar appearance sets.

At present, two types of work are mainly carried out for re-identification tasks. One is to regard the re-identification task as a classification problem, that is, according to the labeled vehicle information as the supervision condition, inputting a large amount of vehicle image data, and using the classification loss function for classification learning. The loss is calculated according to the predicted vehicle category information, and the loss of the classification learning is reduced by continuous forward propagation and backward feedback, thereby realizing the fine-grain classification task of the vehicle. However, the number of vehicle models that appear in the traffic monitoring video is large, and the types of models and the number of vehicles increased year by year. Therefore, using classification learning, that is, regard the re-identification as a fine-grained vehicle classification task, will lead to over-fitting in the data domain. When there are many samples, it will be difficult to classify learning effectively, which leads to bottlenecks in accuracy.

Another type of work carried out for vehicle re-identification tasks is the vehicle verification problem, that is, inputting two vehicle pictures marked with the vehicle ID information, and determining whether the two vehicles belong to the same ID. By using verification loss for verification learning, the loss is reduced gradually to meet the requirements for the distinction between the two vehicles. However, verification learning can only judge the similarity of two pictures in pairs, but because it is a one-to-one comparison, it takes a long time, so it is difficult to apply to target clustering and retrieval. Besides, the generalization ability and the representation ability of the verified verification model are insufficient only by the ID information of the vehicle. Therefore, it is necessary to introduce vehicle attribute labels, such as model and color, and enhance the learning ability and representation ability of the validation model by ''feeding'' enough labeled information into the neural network.

In summary, representation ability plays an important role in vehicle re-identification, methods based on representation learning can automatically extract target features according to task requirements, besides, they are relatively robust, training is more stable and the results are easily reproducible, however, methods based on representation have poor generalization ability, they are easy to over-fitting on the dataset domain, and they appears to be weak when the number of training samples increases to a certain extent.

3) VEHICLE RE-IDENTIFICATION METHODS BASED ON METRIC LEARNING

Metric learning [53], that is, distance metric learning or similarity learning, is a method of mapping into feature space by feature transformation and then forming clusters in feature space. Methods based on metric learning are widely used for face recognition, person re-identification, and vehicle re-identification. Metric learning learns the similarity of two images through the network so that the distance of similar targets becomes closer, and the distance of different targets becomes farther. Taking vehicle re-identification as an example, the metric learning makes the distance between two vehicles belonging to the same ID smaller than the different IDs. (the similarity between vehicles belonging to the same ID is high, and the similarity between vehicles belonging to different IDs is low). Therefore, metric learning requires certain key features of the learning objectives, that is, individualized features. When distinguishing different vehicles, the appearance characteristics of the vehicles are very similar, these features belong to the common features between vehicles. Distinguishing features like the paint, stickers, scratch marks on the vehicle, the annual inspection position of the vehicle on the front windshield, decoration, and tissue boxes are used to distinguish the different characteristics of the two cars. Metric learning distinguishes different identities by learning key distinguishing features.

Commonly used methods of metric learning loss include contrastive loss, triple loss, quaternion loss, etc. Enter two pictures X_1 and X_2, and the feature vectors f_x1, f_x2 can be extracted through the forward propagation of the network. Using Euclidean distance to characterize similarity, define the

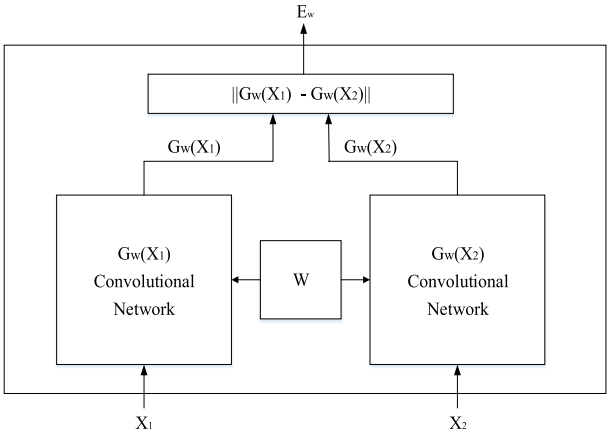

FIGURE 3. Schematic diagram of the siamese network structure.

Euclidean distance formula as (1):

$$D_{x1,x2} = \|f_{x1} - f_{x2}\|_2 \tag{1}$$

a: CONTRASTIVE LOSS

The contrastive loss is used to train the Siamese Network. The Siamese Network is a ''connected neural network'', and its network structure is shown in Fig. 3. The ''connected body'' of the neural network is realized by sharing the weight, that is, the weights of the two neural networks are the same. The Siamese Network is mainly used to measure the similarity between two inputs, both sides can be CNN or LSTM. For example, when two images are input, the two inputs are fed into two neural networks, these two neural networks map the inputs to the new space separately, allowing the input to be represented in the new space. The similarity of the two inputs is evaluated by calculating the loss value.

Taking the vehicle re-identification as an example, the input of the Siamese Network is a pair of vehicle pictures X_1 and X_2, which may be vehicles belonging to the same ID or vehicles belonging to different IDs. Each pair of training pictures has a label y, where $y = 1$ means that the two pictures belong to the same ID, that is, positive sample pairs; $y = 0$ means that two pictures belong to different IDs, that is, negative sample pairs. Contrastive loss function is shown as (2).

$$L_c = \frac{1}{2N}\sum_{i=1}^{n} y d_{X_1,X_2}^2 + \sum (1-y) max\left(0, margin - d_{X_1,X_2}\right)^2 \tag{2}$$

where $d_{X_1,X_2} = \left|\left|f_{x_1} - f_{x_2 1}\right|\right|_2$ represents the calculation of Euclidean distance after feature extraction of two inputs, i.e. the similarity. As mentioned above, y is the label for which two samples match or not, $y = 1$ means that the two samples are similar or matched, $y = 0$ means the two samples are dissimilar or not matched, *margin* is the set threshold. It can be known from the expression of contrastive loss that the loss function can be well expressed for matching between pairs of samples, and can also be effectively used to train models for extracting features.

When $y = 1$, the loss function is $Loss = \frac{1}{2N}\sum_{i=1}^{n} y d_{X_1,X_2}^2$, if the value of d is large, it means that the two samples have a large distance in the feature space, if the two samples are not similar, it is accordant with the require of model training. However, if the two samples are similar, it should be a small distance, but the Euclidean distance in the feature space is large, indicating that the current model is not effective, so the value of loss is larger. To reduce the loss value, the network needs to continue learning.

When $y = 0$, the loss function is $Loss = \sum max\left(0, margin - d_{X_1,X_2}\right)^2$, If the two samples are not similar, the Euclidean distance under ideal conditions is large, but if the Euclidean distance of the feature space is small, the loss value will become larger, which also indicates that the model is not effective and needs continue learning. Through the continuous reduction of the loss value, the distance between the similar sample pairs is continuously reduced, and the distance between the dissimilar sample pairs is continuously increased. In this way, vehicles of different IDs will be distinguished.

Many vehicle re-identification methods are based on the Siamese Network. Zakria *et al.* [10] proposed a novel vehicle re-identification approach, first they chose the vehicle from a gallery set according to appearance, and then verified the chosen vehicle's license plates with a query image to identify the targeted vehicle. In the model, the global channel extracted the feature vector from the whole vehicle image, and the local region channel extracted more discriminative and salient features from different regions. In addition to this, they jointly incorporated attributes like model, type, and color, and Siamese neural network was used to verify the accuracy of re-identification. Liu *et al.* [54] proposed PROVID which used a step-by-step method to search for vehicles (from coarse to fine, from near to far). The appearance properties (color, texture, shape, type) model learned by deep neural network is used as a coarse classifier. The license plate image was matched according to the license plate based on the Siamese Network, and the search process was assisted according to the relationship of time and space, they reordered the vehicles and got the result.

Some jobs combined Siamese Network and other methods to realize vehicle re-identification. Shen *et al.* [55] proposed a two-level framework that contained Siamese-CNN network and Path-LSTM model, one branch network used Siamese Network to calculate visual similarity, and another branch network calculated space-time similarity, which would merge spatio-temporal information converged into the re-identification results. More specifically, firstly, a series of candidate visual-spatio-temporal paths with the query images as the starting and ending states were found. Then, the proposed framework was utilized to determine whether each query pair has the same identity with the spatio-temporal regularization from the candidate path, all the visual-spatio-temporal states were incorporated to estimate the validness confidence of the path. Cui *et al.* [56] proposed a vehicle re-identification method based on deep learning which

exploited a two-branch Multi-DNN Fusion Siamese Neural Network (MFSNN), the MFSNN fused the classification outputs of color, model and pasted marks on the windshield and mapped them into a Euclidean space where distance could be directly used to measure the similarity of arbitrary two vehicles.

Besides, Zhu *et al.* [57] proposed a joint feature and similarity deep learning (JFSDL) method which applied a Siamese deep network learned under the joint identification and verification supervision to extract deep learning features for an input vehicle image pair simultaneously. The joint identification and verification supervision were realized by linearly combining two softmax functions and one hybrid similarity learning function that provide a stronger similarity measurement ability. The advantage of this method was to better explore the identification and verification supervision for training a deep learning-based vehicle re-identification model. Liu *et al.* [58] proposed PROVID, a PROgressive vehicle re-identification framework based on deep neural networks, their framework not only utilized the multimodality data in large-scale video surveillance, such as visual features and contextual information, but also considered vehicle re-identification in two progressive procedures: coarse-to-fine search in the feature domain, and near-to-distant search in the physical space, they adopted a Siamese neural network to verify license number plates for precise vehicle search. Zhu *et al.* [59] proposed a shortly and densely connected convolutional neural network (SDC-CNN) for vehicle re-identification, the SDC-CNN applied a siamese architecture, which included two parameters shared deep feature learning branches and effectively improved the ability of feature learning.

b: TRIPLET LOSS

Compared to the contrastive loss, the input of the triplet loss is changed from two inputs to three inputs, and the network structure is as shown in Fig. 4. The three inputs are an anchor (abbreviated as a), a positive sample belonging to the same ID as the anchor (abbreviated to as p), and a negative sample belonging to different IDs as the anchor (abbreviated to as n). Where a and p are positive sample pairs, a and n are negative sample pairs, and the triplet loss function formula is shown as (3):

$$L_t = (d_{a,p} + \alpha - d_{a,n})_+ \tag{3}$$

Inter-class similarity and intra-class differences are the two basic problems of re-identification tasks. To solve these problems, many frontier methods [60]–[62] use deep networks to learn feature embedding spaces to maximize inter-class distances while minimizing the distance within the class. Schroff *et al.* [60] explored the topic of metric learning to perform k-nearest neighbor classification and proposed the Large Margin Nearest Neighbor loss (LMNN). FaceNet [61] used a modified triplet loss to improve the LMNN loss, the modified triplet loss was used to learn feature embedding based on the principle that ''samples belonging to the same vehicle ID are closer than samples belonging to different IDs'', the final distance metric was directly optimized due to the modified triplet loss, and this triple loss had been widely used for person re-identification and face recognition tasks. Based on the triples, Chen *et al.* [63] also proposed a quadruple network to improve the generalization ability of feature representation. Yang *et al.* [64] used privileged information and unlabeled samples as auxiliary data to construct discriminant metrics. In [65], Zhang *et al.* proposed using multiple labels to inject inter-class relationships (different models, brands, manufacturing years, etc.) as prior knowledge into learning feature representations, without studying the effects of intra-class differences in feature distribution. Wen *et al.* [66] proposed to learn the best center of the deep features of each class and punish the distance between the deep features and their corresponding class centers. Some related work is devoted to introducing semantic knowledge into metric learning, Cui *et al.* [67] designed a general knowledge map to capture the conceptual relationships in the image representation, and then used the regular regression model to jointly optimize image representation learning and graphics embedding. Besides, Li and Tang [68] explored how to use user-provided tags to learn distance metrics, which could reflect semantic information and improve the performance of tag-based image retrieval.

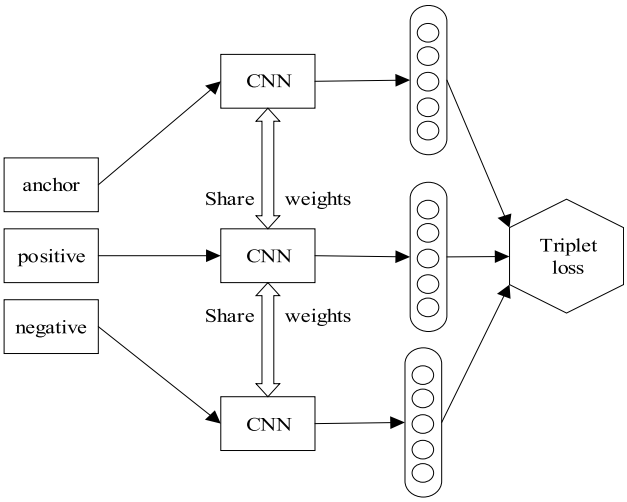

FIGURE 4. Schematic diagram of triplet network structure.

When using triplet loss for metric learning, the goal is to get samples that belong to the same tag (vehicles belonging to the same ID) as close together as possible in the feature space, and other samples that do not belong to the same tag (vehicles belonging to the different IDs) as far as possible, as shown in Fig. 5, images are selected from [92]. Through continuous learning, the vehicle samples of the same ID are finally clustered in the feature space, thereby completing the task of vehicle re-identification.

In terms of vehicle re-identification, inspired by the proposed triplet loss, Liu *et al.* [69] proposed a Deep Relative Distance Learning (DRDL) method which exploited a

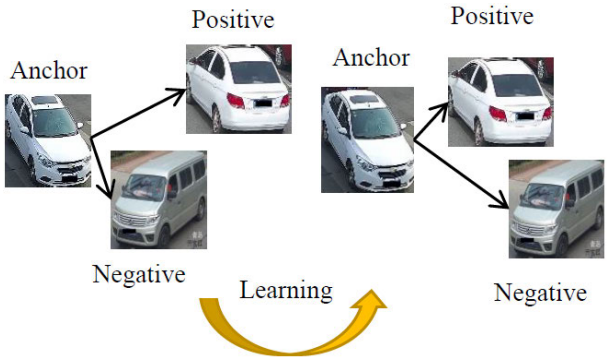

FIGURE 5. Schematic diagram of triplet mapping to feature space learning process.

two-branch deep convolutional network to project raw vehicle images into an Euclidean space where distance could be directly used to measure the similarity of arbitrary two vehicles, the distance could be directly used to measure the similarity of arbitrary two vehicles. Bai *et al.* [70] proposed a deep metric learning method, group-sensitive-triplet embedding (GS-TRE), group sensitive triples were embedded in GSTE, and each ID car was regarded as a category, vehicles with the same ID were divided into the same group, thus, intra-class differences could be better solved for vehicle reidentification and retrieval. Kumar *et al.* [71] solved the problem of vehicle reidentification with utilizing triplet embeddings, and they proposed a detailed evaluation of the contrastive loss function and the triple loss function used in metric learning and proposed a baseline for embedding in the triples for vehicle re-identification under the camera. Zhang *et al.* [72] studied Triplet-wise training which adopted triplets of query, positive example, negative example to capture the relative similarity between them, they proposed a classification-oriented loss that was augmented with the original triplet loss, which essentially improved the traditional triplet loss in enabling stronger classification constraint.

Besides, Li *et al.* [73] proposed a deep joint discriminative learning (DJDL) method to train a convolutional neural network which was aimed to extract discriminative feature representations of vehicle images. DJDL incorporated four different subnetworks in a framework, identification and attribute recognition was to exploit specific properties the individual samples, verification task was to constrain relationship between two samples, and triplet task is responsible for constraining the relative distance among three samples. Finally, an efficient batch composition design was proposed to jointly optimize the four objective functions. Chu *et al.* [74] thought extremely viewpoint variation for vehicles (i.e. 180 degrees) was still very challenging although deep metric learning was useful in getting viewpoint invariant features, they found vehicles with same ID and different views had larger distances than vehicles with different IDs and the same view by experiment, which severely deteriorated the accuracy. Inspired by the human's behavior that a human adopted different strategies when confronted with vehicle images from a similar viewpoint and different viewpoint, they propose a novel viewpoint-aware metric learning approach, named Viewpoint-Aware Network (VANet), learned two metrics for similar viewpoints and different viewpoints in two feature spaces respectively.

For supervised learning, the category is usually fixed, so that the softmax cross-entropy loss function can be used to train to meet the classification requirements. But sometimes, the category is a variable, especially for vehicles, the variety of models is different and will be updated or the quantity will change at any time. The trained classification model has poor generalization ability or is prone to over-fitting, so vehicle re-identification tasks are not well done with only use classification learning, using triplet loss can solve such problems.

In summary, the advantage of triplet loss lies in detail differentiation, that is, when two inputs are similar samples, triplet loss can better model the details and complete the measurement of the different characteristics between the input samples. When distinguishing vehicles, the appearance characteristics of very similar vehicles are regarded as the common features between vehicles and not regarded as the focus of triplet loss, instead, learning differentiated features such as painting, scratch marks on the vehicle, on the front windshield, the vehicle's annual inspection location, decorative objects, tissue boxes, that is, learning a better representation of the input, resulting in a higher accuracy when completing the re-identification work.

4) VEHICLE RE-IDENTIFICATION METHODS BASED ON UNSUPERVISED LEARNING

Most approaches dealing with the re-identification issues are under supervision which affect generalization ability, e.g. training requires a lot of labeled data. While unsupervised learning techniques can potentially cope with such issues by drawing inference directly from the unlabeled input data [75], and have been effectively employed in the context of person re-identification [76]–[78]. Deng *et al.* [77] presented an unsupervised approach for image to image cross domain adaption using the self-similarity and domain-dissimilarity in the training, they used the similarity preserving GANs consisting of the Siamese Neural Networks using the contrastive loss for the re-identification purpose. Wang *et al.* [78] presented a joint attribute-identity learning based approach to simultaneously learned both semantic and attributes in the source domain and transferred it to the target domain to realize unsupervised learning.

Some researchers have applied unsupervised methods to vehicle re-identification. A progressive two step cascaded framework was presented in [75], which essentially formulated the whole vehicle re-identification problem into an unsupervised learning paradigm, it combined a CNN architecture for feature extraction and an unsupervised technique to enable self-paced progressive learning, it also incorporated the contextual information into the proposed progressive framework that significantly improved the convergence

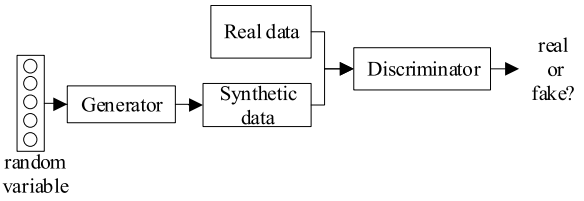

FIGURE 6. Simulative figure of GAN's framework.

of the learned algorithm. Marín-Reyes *et al.* [79] applied the method in [80] to the vehicle re-identification task to create an annotation in an unsupervised manner, along with exploiting visual tracking to produce a weakly labeled training set. Bashir *et al.* [81] presented an unsupervised approach to solve the vehicle re-identification problem by training a base network architecture with a self-paced progressive unsupervised learning architecture, the technique enabled the transfer of deeply learned representation towards unlabeled dataset.

The generative adversarial network (GAN) [82] is an emerging technique for unsupervised learning. It includes a generator and a discriminator. The generator obtains random variable from the prior distribution and obtains synthetic data through the transformation of generator. The discriminator receives both synthetic data from the generator and real data, and it needs to determine the source of data. The generator confuses the discriminator with synthetic samples that are as close to the real data as possible while the discriminator identifies the synthetic data generated by the generator as possible. Ideally, the generator and discriminator finally reach a balance, and both sides tend to be perfect. Simulative figure of GAN's framework is shown in Fig. 6.

GANs have achieved great success on many visual tasks, such as image generation [83], [84] and image translation [85] In essence, the design of confrontational learning led to the success of GANs, mainly because it forced the generated samples to be indistinguishable from the actual data. Besides, many efforts extend GANs to conditional GANs such as InfoGAN [86], AC-GAN [87], and CycleGAN [88] to study generation models with better performance.

The breakthrough of GANs in image generation inspires people to generate vehicles from different viewpoints. In the vehicle re-identification problem, to solve the vehicle re-identification task under multiple viewing angles, a multi-view feature could be generated for each image by GANs given a query vehicle image and a set of gallery images, which can be regarded as containing the descriptive representation of all perspective information [89]. Firstly, extracting features of an input image containing only single-view visual content. Then, obtaining the correlation between the input perspective and other hidden perspectives by modeling, the transformation model was learned to infer features from other perspectives. Finally, all the features in different perspectives were merged, and the feature was used to embed the feature into the distance space on the end-to-end network. The biggest difference of this method is that it was inspired by the GAN to transform the single-view feature into multi-view feature in the form of generation/confrontation, two generators were used in the way, the input of G_f was the attention feature of the single-view image, and the input of G_r was the feature of the real picture of different views with the same ID as G_f. The goal of the generator G_f was not to maximize the output of the discriminator, but to have the same statistical distribution of the single-view data in the fourth layer of the discriminator D having the same layer characteristics as the multi-view data and completing a multi-view feature for a single-view image. Similar to this approach, to augment the training data for robust training, Wu *et al.* [90] adopted a Generative Adversarial Network to generate unlabeled samples and enlarge the training set, besides, a semi-supervised learning scheme with the CNNs was proposed to improve the performance of the vehicle re-identification system, which assigned a uniform label distribution to the unlabeled images to regularize the supervised model. In [91], a generative adversarial network was used to synthesize vehicle images with diverse orientation and appearance variations to obtain more vehicle images and augment the training set.

Many GAN-based methods have been proposed to improve the robustness and accuracy of vehicle re-identification. Lou *et al.* [92] proposed a Feature Distance Adversarial Network (FDA-Net) which aimed to explore generating hard negatives in the feature space to improve the discriminative capability of the re-identification model. It contained a novel adversary scheme on feature distance between the generator G and the embedding discriminator D. The G tried to generate a hard-negative sample under similarity constraint and attention regularization while the D tried to discriminate them, the generator and discriminator were alternatively optimized. An end-to-end embedding adversarial learning network (EALN) was proposed in [93], it could generate samples localized in the embedding space, with its embedding adversarial learning scheme instead of selecting abundant hard negatives from the training set. The automatically generated hard negative samples in the specified embedding space could improve the capability of the network for discriminating similar vehicles. Besides, the model was able to generate desired vehicle images from same-view and cross-view, which facilitated re-identification model training as well as improved the discriminative capability and robustness of the re-identification algorithm. Zhou and Shao [94] proposed Cross-View Generative Adversarial Network (XVGAN) to learn the features of vehicle images captured by cameras with disjoint views. They took the features as conditional variables to effectively infer cross-view images. They combined the features of the original images and the features of generated images in other views to learn distance metrics for vehicle re-identification. The proposed model could successfully generate realistic images in different views of the same vehicle, and improved the accuracy of vehicle re-identification.

There was a method note the inconsistency in the distribution of different data sources. When deploying the well-trained model to a new dataset directly, there is a severe

performance drop because of differences among datasets named domain bias, Peng *et al*. [95] proposed a domain adaptation framework to address this problem, which contained an image-to-image translation network named vehicle transfer generative adversarial network (VTGAN) and an attention-based feature learning network (ATTNet). The advantage of VTGAN is that the source domain (well-marked) image can have the style of the target domain (unmarked) and the source domain identity information can be retained.

In summary, unsupervised learning technology can make use of unmarked input data to improve generalization ability. Among the vehicle re-identification methods based on unsupervised technology, GAN-based methods are widely used. GANs can generate multiple perspective features for a single perspective image and using the feature to solve the vehicle re-identification problem under multiple viewing angles, in addition, GAN can be used for image to image translation to better solve the problem of inconsistent distribution of different data domains. But using GANs for image generation needs to overcome the problem of difficulty in convergence, and balance the two models in training, thereby avoid unstable training situations

5) VEHICLE RE-IDENTIFICATION METHODS BASED ON ATTENTION MECHANISM

In recent years, most researches on the combination of deep learning and visual attention mechanism focused on using masks to form the attention mechanism. The mask works by identifying key features in the image data with another layer of new weight, attention is formed by training deep neural networks to learn what areas need to be focused on in each new image, this idea evolved into two different types of attention, soft attention and hard attention. The key point of soft attention is that it pays more attention to areas [96] or channels [97], and soft attention is deterministic attention, which can be generated directly through the network after learning. The most critical place is that soft attention is differentiable, which is a very important place, differential attention can be used to calculate the gradient through neural network and forward propagation and backward feedback to learn the weight of attention [98]. The difference between strong attention [99] and soft attention lies in that, strong attention is more focused on points, that is, every point in the image is likely to extend the attention. Meanwhile, strong attention is a random prediction process, with more emphasis on dynamic changes.

Attention mechanism has explored in many applications, such as image classification [100], [101], fine-grained image recognition [102], [103], image captioning [104], [105], and VQA [106], a growing number of researchers are using attentional mechanisms in vehicle re-identification. Guo *et al*. [9] proposed a Two-level Attention network supervised by a Multi-grain Ranking loss (TAMR) to learn an efficient feature embedding for vehicle re-identification task, the two-level attention network included hard part-level attention and soft pixel-level attention. Hard part-level attention was designed to localize the salient vehicle parts. Soft pixel-level attention gave an additional attention refinement at pixel level to focus on the distinctive characteristics within each part. Therefore, the two-level attention network could adaptively extract discriminative features from the visual appearance of vehicles. Based on the Region-Aware deep Model [37], Chang *et al*. [107] proposed a Pyramid Granularity Attentive Model (PGAM) such that both coarse and fine-grained features could be effectively extracted, and fine-grained discriminability could be retained by adopting many improved model training approaches.

There are methods based on hard attention. Khorramshahi *et al.* [108] found that the contribution of each key point was different depending on the direction, while most re-identification methods were designed to focus attention at key-point locations, so they presented a dual path adaptive attention model for vehicle re-identification (AAVER), the global appearance path captured macroscopic vehicle features while the orientation conditioned part appearance path learned to capture localized discriminative features by focusing attention to the most informative key-points. Khorramshah *et al.* [109] presented an attention-based model which learned to focus on different parts by conditioning the feature maps on visible key-points. They used different datasets to train networks, and used triplet embedding to reduce the dimensionality of the features obtained from the ensemble of networks.

There is a method based on soft attention. Teng *et al.* [110] proposed a Spatial and Channel Attention Network (SCAN) based on DCNN, the attention model contained a spatial attention branch and a channel attention branch, the two branches adjusted the weights of outputs in different positions and different channels to highlight the outputs in discriminative regions and channels respectively. Feature maps were refined by the attention model and more discriminative features can be extracted automatically.

There are many other methods. When humans identify different vehicles, humans always firstly determined one vehicle's coarse-grained category such as the car type, and then identified specific vehicles by relying on subtle visual cues, such as windshield stickers at the fine-grained level. Inspired by this, Wei *et al.* [111] proposed an end-to-end RNN-based Hierarchical Attention (RNN-HA) classification model for vehicle re-identification. The RNN- HA consisted of three models, the first generated image representations, the second modeled the hierarchical dependent relationship, and the last focused on capturing the subtle visual information to distinguish specific vehicles from each other. Besides, Zhang *et al.* [112] introduced a Part-Guided Attention Network (PGAN), the PGAN combined art-guided bottom-up and top-down attention, global and part visual features in an end-to-end framework. PGAN first detected the locations of different part components and salient regions, which served as the bottom-up attention to narrow down the possible searching regions, a Part Attention Module (PAM) were

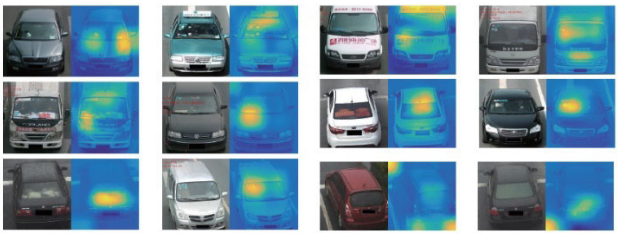

FIGURE 7. Attention maps.

proposed to adaptively locate the most discriminative regions with high-attention weights and suppressed the distraction of irrelevant parts with relatively low weights. The PAM was guided by the re-identification loss and therefore provides top-down attention. Finally, global appearance and part features were aggregated to improve the feature performance further.

In summary, the attention mechanism mimics the process of re-identification of humans, deep neural networks can learn what areas need to be focused on by training. Fig. 7 (from [111]) shows the learned attention maps of some vehicle images, the attended regions accurately correspond to these subtle and discriminative image regions, such as windshield stickers, and customized paintings. The attention mechanism automatically extracts the features of the distinguishing regions, resulting in the improvement of accuracy in the vehicle re-identification task. However, it can be found that most attention-based models focus their attention on regions and pay less attention to the differences of finer pixel level, when the datasets are less labeled and the background is more complex, the method based on attention mechanism is less effective.

6) OTHER VEHICLE RE-IDENTIFICATION METHODS

Apart from the five major types of deep learning-based methods mentioned above, there have been some other deep learning-based researches on vehicle re-identification.

For better handling viewpoint variations, Zhou and Shao [113] proposed the adversarial bidirectional long short-term memory network (ABLN), ABLN used long short-term memory network (LSTM) to model transformations across continuous view variations of a vehicle and adopted the adversarial architecture to enhance training. Zhou *et al.* [114] proposed two end-to-end deep architectures: a spatially Concatenated ConvNet and a CNN-LSTM bi-directional loop, which exploited the great advantages of the CNN and LSTM to learn transformations across different viewpoints of vehicles. To model a view-invariant similarity between vehicle images from different views, Zheng *et al.* [115] proposed a Ranked Semantic Sampling (RSS) guided binary embedding method for fast cross-view vehicle Re-identification. Vehicle re-identification problem was modeled as two sub tasks in [116], including the same view and across different views, a fine-grain ranking loss and a relative coarse-grain ranking loss were proposed to each task respectively. Xu *et al.* [117] presented a multi-scale vehicle re-identification framework using self-adapting label smoothing regularization (SLSR), it integrated the appearance information from multi-scale images to alleviate the influence of scale changes caused by perspectives, besides, self-adapting label smoothing regulation was designed in the semi-supervised training process to enhance the generalization ability. Zhu *et al.* [118] proposed a quadruple directional deep learning which utilized different directional pooling layers to compress the basic feature maps into horizontal, vertical, diagonal and anti-diagonal directional feature maps, respectively, and then spatially normalized these directional feature maps and concatenated together as a quadruple directional deep learning feature for vehicle re-identification which improved the robustness of viewpoint variations. Zhu *et al.* [119] proposed the joint horizontal and vertical deep learning feature (JHV-DLF), it aimed to describe vehicle images in both horizontal and vertical directions and makes re-identification robust toward view-point variations. Tang *et al.* [120] proposed a Pose-Aware Multi-Task Re-Identification (PAMTRI) framework, which overcame viewpoint-dependency by explicitly reasoning about vehicle pose and shape via keypoints, heatmaps, and segments from pose estimation, it jointly classified semantic vehicle attributes (colors and types) while performing re-identification, through multi-task learning with the embedded pose representations. Since manually marking images with detailed attitude and attribute information is time-consuming and labor-intensive, they create a large-scale highly randomized synthetic dataset with automatically annotated vehicle attributes for training.

There are other methods, Liang *et al.* [121] proposed a new supervised deep hashing method to deal with large-scale instance-level vehicle search, which utilized sigmoid or tanh as the activation function of the hash layer, rectified linear unit and showed better performance. To fully explore the complementary correlation between learning-based deep features and hand-crafted features, Tang *et al.* [122] proposed a multi-modal metric learning architecture, which fused deep features and handcrafted ones in an end-to-end optimization network, which achieved a more robust and discriminative feature representation for vehicle re-identification. Hou *et al.* [123] proposed a deep quadruplet appearance learning (DQAL), which lied on the consideration of the special difficulty in vehicle re-identification that the vehicles with the same model and color but different IDs are highly similar to each other, each quadruplet in DQAL was composed of the anchor, positive, negative, and the specially considered high-similar vehicle samples, quadruplet loss and softmax loss was developed to learn a more discriminative feature. Besides, Huang *et al.* [124] proposed a viewpoint-aware temporal attention model for vehicle re-identification utilizing deep learning features extracted from consecutive frames with vehicle orientation and metadata attributes (i.e., brand, color) being taken into consideration. Kanaci *et al.* [125] proposed a novel Multi-Task Mutual Learning (MTML) deep model to learn discriminative

TABLE 2. Multi-dimensional comparison of vehicle re-identification algorithms based on deep learning.

Method	Characteristics	Advantages	Disadvantages
Local Feature	Key point location and region segmentation	1. Capture unique visual clues 2. Improve the perception of nuance	Increase the computational burden
Representation Learning	Focus on vehicle attribute characteristics	1. Easy to train 2. Training is stable 3. Relatively robust	1. Poor generalization ability 2. Easy to over-fitting on the dataset domain
Metric Learning	Focus on details of vehicle	High accuracy	Training is unstable and difficult to converg
Unsupervised Learning	No need of labeled information	1. Solve the effect of perspective change 2. Improve generalization ability 3. Solve the problem of inconsistent distribution of different data domains	Unstable training
Attention Mechanism	Self-adaptive extract features	1. Learn what areas need to be focused on by training 2. Extracts the features of the distinguishing regions	Poor effect when few labeled data and complex background

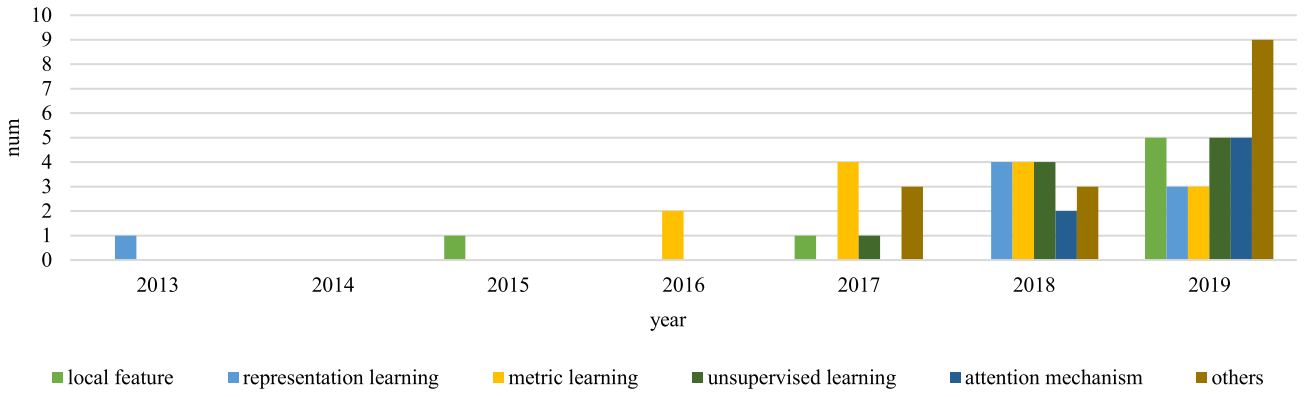

FIGURE 8. Number of vehicle re-identification papers based on deep learning for each category from 2013 to 2019.

features simultaneously from multiple branches, specifically, they designed a consensus learning loss function by fusing features from the final convolutional feature maps from all branches. Kanaci and Gong [126] proposed Cross-Level Vehicle Recognition (CLVR), they transferred the vehicle model discriminative representation for more fine-grained re-identification tasks by fully leveraging the strong capacity of existing deep models in learning cross-level representations. Vehicle re-identification suffers from varying image quality and challenging visual appearance characteristics, training CNNs on multiple datasets simultaneously is a solution to enhance the feature robustness, however, due to misaligned feature distribution between domains, the larger set of training data does not guarantee performance improvement, Liu *et al.* [127] proposed a Joint Domain Re-Identification Network (JDRN) to mitigate the domain gap, which improved the feature by disentangling domain-invariant information and encouraged a shared feature space between domains.

C. SUMMARY

In order to further develop the vehicle re-identification methods based on deep learning, this paper compares the vehicle re-identification method from characteristics, advantages and disadvantages, as shown in Table 2; The papers about vehicle re-identification method based on deep learning in 2013-2019 were classified, and we summarize the characteristics of each method, as shown in Table 3. We count the number of vehicle re-identification papers based on deep learning for each category from 2013 to 2019, as shown in Fig. 8. There were few vehicle re-identification methods based on deep learning from 2013 to 2016, and more methods are based on sensors and traditional machine learning. From 2017 to 2019, the number of vehicle re-identification methods based on deep learning gradually increased, indicating that with the development of deep learning, methods deep learning could be better used to solve vehicle re-identification problems. In terms of the number of

TABLE 3. Classified statistics of papers about vehicle re-identification methods based on deep learning in 2013-2019.

Category	Method and Paper	The characteristics of the method
Local Features	OIFE [36]	Compared different regions between different vehicle images
	RAM [37]	Embed the detailed visual cues in local regions
	Part-regularized Near-duplicate [38]	Enhanced the perceptive ability of subtle discrepancies
	MRM [39]	Localize local regions that contain more distinctive visual cue
	PRN [40]	Extracted more local features from height, width and channel
	ROIs-based vehicle re-identification [41]	Local features of ROIs could be extracted by the location of SSD detector
	A refined part model[42]	Automatically locate the vehicle and perform division for local features
Representation Learning	DF-CVTC [45]	Jointly learn deep feature representations (views, types, colors)
	DFFMG [46]	Integrated discriminative information with various granularity
	Random occlusion assisted deep representation learning [47]	Simulated some occlusion situations in real world
	3D-OR[48]	Lifted two-dimensional object representations to three-dimensional representations
	Mob.VFL[49]	Generate variational features which were more discriminating
	Space-Time Prior [50]	Address issues like visual domain mismatch between datasets
	A new distance loss [51]	Accelerate the procedure of feature learning
	Multi-attribute driven vehicle re-identification [52]	Explicitly leveraged the vehicle attribute cues such as color, model
Metric Learning	Multi-Level Feature Extraction [10]	Jointly incorporated attributes like model, type, and color
	PROVID [54]	Search for vehicles from coarse to fine, from near to far
	Path-LSTM [55]	Merge spatio-temporal information
	MFSNN [56]	Fused the classification outputs of color, model, pasted marks on the windshield
	JFSDL [57]	Extract deep learning features for an input vehicle image pair simultaneously
	PROVID [58]	Utilized the multimodality data in large-scale video surveillance
	SDC-CNN [59]	Effectively improved the feature learning ability
	DRDL[69]	Directly used to measure the similarity of arbitrary two vehicles
	GSTE [70]	Intra-class differences could be better solved for vehicle reidentification
	Batch sample [71]	Proposed a baseline for embedding in the triples for vehicle re-identification
	ITWT-CNN [72]	Improved the traditional triplet loss in enabling stronger classification constraint
	DJDL [73]	Extract discriminative feature representations of vehicle images
	VANet[74]	Learned two metrics for similar viewpoints and different viewpoints
Unsupervised Learning	VR-PROUD [75]	A CNN and an unsupervised technique to enable self-paced progressive learning
	UVRTN [79]	Create an annotation in an unsupervised manner
	Deep Unsupervised Progressive Learning [81]	Training a base network with a self-paced progressive unsupervised learning.
	VAMI [89]	A multi-view feature could be generated for each image by GANs
	GAN[90]	Generate unlabeled samples and enlarge the training set
	Semi-supervised learning and re-ranking [91]	Synthesize vehicle images with diverse orientation and appearance variations
	FDA-Net [92]	Explore generating hard negatives in the feature space
	EALN [93]	Generate desired vehicle images from same-view and cross-view
	XVGAN[94]	Effectively infer cross-view images by learning the features of vehicle images
	VTGAN [95]	Make images from the source domain have the style of target domain
	TAMR [9]	Learn an efficient feature embedding for vehicle re-identification
	PGAM [107]	Both coarse and fine-grained features could be effectively extracted

TABLE 3. *(Continued.)* **Classified statistics of papers about vehicle re-identification methods based on deep learning in 2013-2019.**

Attention Mechanism	AAVER [108]	Global appearance path and part appearance path for adaptive attention model
	Attention driven vehicle re-identification [109]	Focused on different parts by conditioning the feature maps on visible key points
	SCAN [110]	More discriminative features can be extracted automatically
	RNN-HA [111]	Use three models from coarse-grained category to subtle visual cues
	PGAN [112]	Combing part-guided bottom-up and top-down attention
Others	ABLN [113]	Use LSTM to model transformations across continuous view variations
	Two end-to-end deep architectures[114]	Exploited the great advantages of the CNN and LSTM to learn transformations across different viewpoints of vehicles
	RSS [115]	Fast cross-view vehicle re-identification
	MRL+Softmax Loss [116]	Two sub tasks including the same view and across different views
	SLSR [117]	Using self-adapting label smoothing regularization
	A quadruple directional deep learning [118]	Utilized different directional pooling layers to compress the basic feature map
	JHV-DLF [119]	Describe vehicle images in both horizontal and vertical directions
	PAMTRI [120]	Overcame viewpoint-dependency
	A new supervised deep hashing method [121]	Utilized sigmoid or tanh as the activation function of the hash layer
	A multi-modal metric learning architecture [122]	Fused deep features and handcrafted features
	DQAL [123]	Learn a more discriminative feature especially for difficult high-similar cases
	A viewpoint-aware temporal attention model [124]	Use features extracted from vehicle orientation and metadata attributes
	MTML [125]	Learn discriminative features simultaneously from multiple branches
	CLVR [126]	Leverage the capacity of deep models in learning cross-level representations
	JDRN [127]	Improved the problem of misaligned feature distribution between domains

methods, there are few methods based on local features and representation learning, while there are many methods based on metric learning. In the last two years, the methods based on unsupervised learning and attention mechanism have developed rapidly, and most of methods based on unsupervised learning are based on GANs. Methods based on unsupervised learning can make use of unmarked input data to improve generalization ability, and GANs can be used to generate multi-perspective features, which is conducive to solving the problem of vehicle re-identification under multi-perspective. Methods based on the attention mechanism can automatically extract the distinguishing features and improve the accuracy of re-identification. Due to these advantages, these two kinds of methods develop rapidly in recent years.

III. DATASET AND EVALUATION STRATEGY

Through the above review of the vehicle re-identification method, the research work in the field of vehicle re-identification mainly focuses on two aspects, one is the classification model and the other is the re-identification algorithm. Datasets are an important prerequisite for classification and re-identification tasks in the field of computer vision. To some extent, the size and characteristics of datasets limit the advancement of research work in this field, and restrict the performance of the classification model or re-identification in some aspects, such as limiting the accuracy or generalization ability. In theory, more training data results in better re-identification effect. Therefore, this chapter will introduce vehicle public datasets and comparing them from multiple dimensions.

A. DATASET

With the increasing demand for accurate identification of vehicle retrieval, vehicle re-identification, vehicle positioning, vehicle tracking, and the deep maturity of deep learning related technologies, the accuracy of vehicle re-identification is increasing. In addition to the fact that today's technology breaks the shackles of hardware performance on re-identification, it also benefits from the expansion of the vehicle public datasets. The most commonly used public datasets for verifying the performance of vehicle re-identification algorithms are VeRi-776 [54], PKU-VD [130], etc. More details about vehicle public datasets will be presented separately.

The BoxCars dataset, including BoxCars21k [128] and BoxCars116K [129] is collected by the Bruno University of Technology in the Czech Republic. BoxCars21K contains 21,250 cars, 63,750 pictures, 27 different brands and 148 models. BoxCars116K contains 27,496 cars, 116,826 pictures, 45 different brands and 693 models. The dataset contains vehicle images captured from arbitrary viewpoints, front, side, and roof. Compared to other fine-grained

FIGURE 9. Sample images from the Boxcars dataset.

FIGURE 10. Example of the VehicleID dataset.

surveillance datasets, the dataset provides data with a high variation of viewpoints. All images are labeled with the 3D bounding box, make, model, and type. The dataset is designed for fine-grained vehicle model, make classification, and re-identification, it can be also be used for vehicle re-identification problem. Sample images from the Boxcars dataset are shown in Fig. 9.

The VehicleID [69] dataset is built by the National Engineering Laboratory for Video Technology of Peking University (NELVT) and sponsored by the China National Basic Research Program and the National Natural Science Foundation of China. The VehicleID dataset contains data captured during the day by multiple real surveillance cameras distributed in a small city in China. There are 26,267 vehicles in the entire dataset (221,763 images in total). Each image is accompanied by an ID tag that corresponds to its identity in the real world. Besides, 10,319 vehicles (90,196 images in total) marked with vehicle model information. Sample images from the VehicleID dataset are shown in Fig. 10.

The PKU-Vehicle [70] dataset is a vehicle dataset for a large-scale real-world scenario proposed by the Peking University team to meet the needs of large-scale vehicle re-identification. It contains images of tens of millions of vehicles taken by real surveillance cameras in several Chinese cities. There are ten million vehicle images which is primarily used as an interference dataset for simulating real-world retrieval. The dataset contains various locations, weather conditions (e.g. sunny, rainy, foggy), lighting (e.g. day and evening), shooting angle and hundreds of vehicle brands and other information. Collecting images of the PKU-Vehicle dataset from different cameras makes the original resolution of the vehicle image vary greatly, which brings more difficulties to the vehicle re-identification. Sample images from the PKU-Vehicle dataset are shown in Fig. 11.

FIGURE 11. Sample images from the PKU-Vehicle dataset.

FIGURE 12. Sample images from the PKU-VD dataset.

The PKU-VD [130] dataset, including VD1 and VD2, is constructed by the National Video Technology Engineering Laboratory (NELVT) of Peking University. Two large vehicle datasets (VD1 and VD2) are constructed based on real-world unconstrained scenes from two cities. The images in VD1 are from high-resolution traffic cameras, and the images in VD2 are captured from surveillance video. Performing vehicle detection from raw data to ensure that each image contains only one vehicle. Each image in both datasets provides different attribute annotations, including identity numbers, precise vehicle models, and vehicle colors. Particularly, the ID is unique and images contain the same vehicle have the same ID. The dataset has the most accurate model of vehicle and vehicles in different production years. As for color information, the dataset is labeled with 11 common colors. To ensure the consistency of the labels, all images belonging to the same vehicle ID are marked with the same model and color. With 1,097,649 and 807,260 images being collected and carefully annotated, the datasets contain almost all popular vehicle models and colors which makes the dataset expandable enough for vehicle re-identification and other related research. Sample images from the PKU-VD dataset are shown in Fig. 12.

VeRi-776 [54] is a public vehicle dataset published by the Beijing University of Posts and Telecommunications on ECCV, it is built from the VeRi dataset [131]. It uses images captured in a real-world unconstrained surveillance scene and labeled images with different attributes, for example, license plate bounding box, models, colors and brands, and it can be used to do vehicle re-identification work. Each car is photographed by 2 to 18 cameras at different viewpoints, illumination, resolution, and occlusion conditions, providing a high recurrence rate for the real situation of the vehicle's re-identification. Besides, it is labeled with enough license

FIGURE 13. Sample images from the VeRi-776 dataset.

FIGURE 14. Sample images from the CompCars dataset.

plates and time-space information, such as the bounding box of the license plate, the time stamp of the vehicle and the distance between adjacent cameras. Sample images from VeRi-776 dataset are shown in Fig. 13.

The CompCars [132] dataset is presented in the CVPR 2015 paper by the Tang's team, it is a large-scale automotive dataset for fine-grained classification and validation which contains data from two scenarios, images from the network and monitoring. The network's data includes 163 cars and 1,716 models, a total of 136,726 images captured the entire car and 27,618 images captured the car parts. The complete car image is marked with a bounding box and a viewpoint. Each model is marked with five attributes including maximum speed, displacement, number of doors, number of seats and type of car. The monitoring data contains 50,000 car images captured in the front view. This dataset can be used for a variety of computer vision tasks: fine-grained classification, attribute prediction, and vehicle model validation. Sample images from the CompCars dataset are shown in Fig. 14.

The Vehicle-1M [133] dataset is constructed by the National Pattern Recognition Laboratory (NLPR, CASIA) of the Institute of Automation, University of Chinese Academy of Sciences. The dataset relates to images of vehicles captured from the head or the back of the night through multiple surveillance cameras installed in several cities in China. There are 936,051 images in 55,527 vehicles and 400 models in the dataset, and they extract a small, medium and large test set from the original test set. Each image is accompanied by a vehicle ID tag indicating its identity in the real world and a vehicle model tag indicating the brand, model and year of the vehicle (e.g. "Audi-A6-2013"). The difference between vehicle models can be quite small, just like the real-world vehicle re-identification situation, thus this dataset is very suitable for vehicle re-identification. Sample images from the Vehicle-1M dataset are shown in Fig. 15.

FIGURE 15. Sample images from the Vehicle-1M dataset.

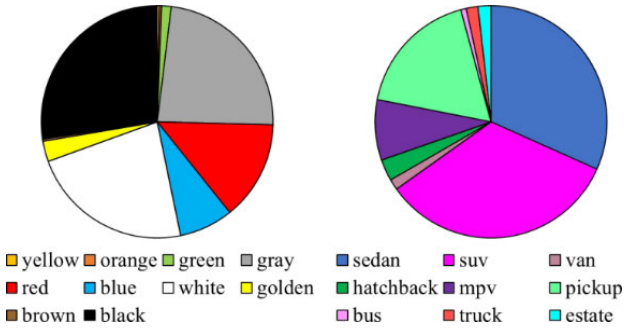

FIGURE 16. Type and color distribution of the CityFlow.

CityFlow [134], the world's first large dataset to support cross-camera car tracking and re-identification, includes 3.25 hours of synchronized HD video collected from 10 intersections and 40 cameras, the longest distance between two sync cameras is 2.5 kilometers. The dataset is collected in a medium-sized US city with a variety of scenarios, including residential areas and highways. The dataset has 229,680 bounding boxes of 666 vehicle identities labeled, each car passes through at least 2 cameras, providing raw video, camera distribution, and multi-view analysis. Type and color distribution of the CityFlow dataset are shown in Fig. 16.

VERI-Wild [92], which is currently the most challenging dataset for vehicle re-identification in real scenarios and the first vehicle re-identification dataset that is collected from unconstrained conditions. It is captured via a large Closed Circuit Television (CCTV) system. It contains 174 surveillance cameras and covers a large urban district of more than 200km2, the 174 cameras capture for 24 hours for 30 days so that various weather and illumination conditions are considered, such as rainy, foggy, etc. The dataset contains 416,314 images of 40,671 IDs after cleaning from 12 million vehicle images. The dataset poses many more practical challenges for vehicle re-identification, such as different viewpoints, illumination, and background variations, and severe occlusion. Sample images from VERI-Wild dataset are shown in Fig. 17. Type and color distribution of the VERI-Wild dataset are shown in Fig. 18.

VRID-1 [135] contains 10,000 images captured in the daytime of 1,000 individual vehicles of the ten most common vehicle models. For each vehicle model, there are

FIGURE 17. Sample images from the VERI-Wild dataset.

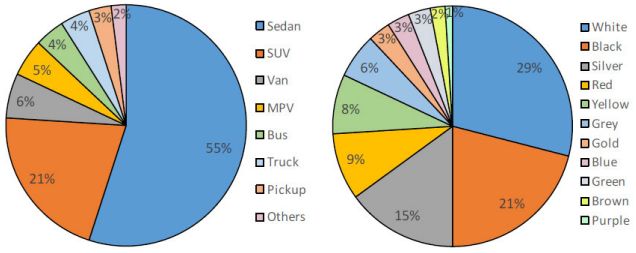

FIGURE 18. Type and color distribution of the VERI-Wild.

100 individual vehicles, and for each of these, there are ten images captured at different locations. The images in VRID-1 were captured by 326 surveillance cameras, and thus there are various vehicle poses and levels of illumination, it provides images of good enough quality for the evaluation of vehicle re-identification in a practical surveillance environment. Sample images from the VRID-1 dataset are shown in Fig. 19.

Toy Car [114] is the first synthetic vehicle dataset collected in an indoor environment using multiple cameras. It is a toy car dataset that contains many common vehicle types such as sedan, SUV, van, and pickup, it consists of 200 different models of toy cars. To reduce the gap in appearance between toy cars and real cars, metal toy cars as real as possible to construct the dataset were selected, lighting is provided to simulate illumination by the sun. collecting sequences of vehicles as they rotated by 360 degrees using a rotation stage, setting cameras at three angles: 30, 60 and 90 to capture data with different altitudes, averagely sampled 50 viewpoints and cropped all the vehicles to generate the raw dataset containing 30,000 images in total, then replacing the green background with random road patterns to synthesize the final toy car dataset. Sample images from the Toy Car dataset are shown in Fig. 20.

VRIC (Vehicle Re-Identification in Context) [136] is a more realistic and challenging vehicle re-identification benchmark, in contrast to other vehicle re-identification datasets, VRIC is uniquely characterized by vehicle images subject to more realistic and unconstrained variations in resolution (scale), motion blur, illumination, occlusion, and viewpoint. It contains 60,430 images of 5,622 vehicle identities captured by 60 different cameras at heterogeneous road traffic

FIGURE 19. Sample images from the VRID-1 dataset.

FIGURE 20. Sample images from the Toy Car dataset.

FIGURE 21. Sample images from the VRIC dataset.

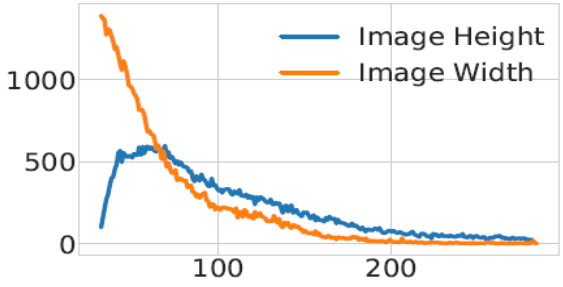

FIGURE 22. Vehicle instance scale distributions in VRIC.

scenes in both day-time and night-time. Sample images from the VRIC dataset are shown in Fig. 21. Vehicle instance scale distributions in VRIC is shown in Fig. 22.

Wang *et al.* [137] construct a large-scale dataset for vehicle re-identification named Vehicle Re-identification for Aerial Image (VRAI), which contains 137,613 images of 13,022 vehicles captured by UAV-mounted cameras. The images of each vehicle instance are captured by cameras of two DJI consumer UAVs at different locations, with a variety of view-angles and flight-altitudes. To increase intra-class

TABLE 4. Multi-dimensional comparison table of vehicle public datasets.

Name	Scale	Number of vehicles	Number of models	Number of colors	Characteristics	Application
BoxCars21k[128]	63,750	21,250	148	/	Monitor	Fine-grained vehicle identification, vehicle re-identification
BoxCars116K[129]	116,826	27,496	693	/	Monitor	
VehicleID [69]	221,763	26,267	250	7	Monitor	Vehicle re-identification
PKU-Vehicle[70]	10,000,000	/	/	/	Monitor	Vehicle retrieval, vehicle re-identification
PKU-VD1 [130]	1,097,649	/	1,232	11	Monitor	Fine-grained vehicle identification, vehicle re-identification
PKU-VD2 [130]	807,260	/	1,112	11	Monitor	
VeRi-776[54]	50,000	776	/	10	Monitor	Vehicle re-identification
CompCars[132]	136,726+27,618	/	1,716	11	Internet Monitor	Vehicle identification
Vehicle-1M[133]	936,051	55,527	400	/	Monitor	Vehicle model categorization, vehicle model verification, vehicle re-identification
CityFlow[134]	229,680	/	/	/	Monitor	Vehicle re-identification, cross-camera vehicle tracking
VERI-Wild[92]	416,314	40,671	/	11	Monitor	Vehicle re-identification, cross-camera vehicle tracking
VRID-1 [135]	10,000	1,000	10	/	Monitor	Vehicle re-identification
Toy Car [114]	30,000	150	200	/	Synthetic dataset	Vehicle re-identification
VRIC[136]	60,430	5,622	/	/	Monitor	Vehicle re-identification
VRA[137]	137, 613	13, 022	/	/	Unmanned Aerial Vehicles	Vehicle re-identification

FIGURE 23. Sample images from the VRAI dataset.

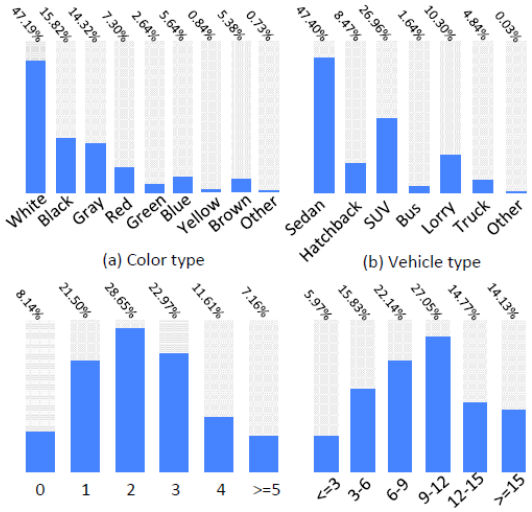

FIGURE 24. Statistical information of the VRAI dataset.

variation, each vehicle is captured by at least two UAVs at different locations, with diverse view-angles and flight-altitudes. There are many well-labeled vehicle attributes, including vehicle type, color, skylight, bumper, spare tire, and luggage rack. Besides, the annotators mark the discriminative parts which is helpful to distinguish a vehicle from others for each vehicle image. The VRAI datasets bring more challenges for vehicle re-identification as vehicles in VRAI are featured in larger pose variation and wider range of resolution. Sample images from the VRAI dataset are shown in Fig. 23. The statistical information about color, vehicle type, discriminative part number per image, and image number per vehicle VRAI are shown in Fig. 24.

The emergence of many datasets used for vehicle re-identification is conducive to the development of re-identification methods, as well as bringing challenges. Through the above detailed description of the data related to the vehicle open dataset, this paper summarizes the vehicle datasets in the field of vehicle re-identification in recent years, and give a list of the scale, number of vehicles, number of models, number of colors, characteristics, and application for each dataset, as shown in Table 4.

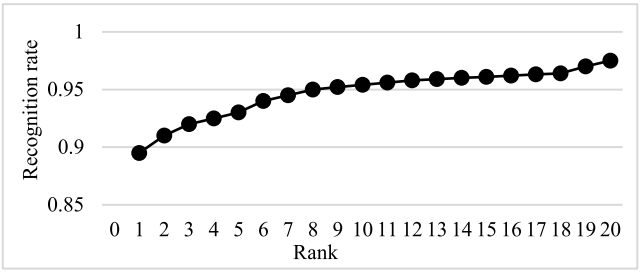

FIGURE 25. An example of CMC Curve.

B. EVALUATION STRATEGY

To measure the performance of the vehicle re-identification model, the researchers proposed some measurement indexes. In this section, some common strategies used to evaluate the performance of the vehicle re-identification algorithm will be introduced.

1) RANK

Rank measures the similarity of a test to its owned class [75]. The value of rank-m is the probability of correct results in images with the highest confidence in the search results, a higher Rank indicates better performance of the model. For example, a car labeled c1 is searched through 100 samples. If the result is c1, c2, c3, c4, c5..., the accuracy rate of rank-1 is 100%, because c1 ranks in the first position of the result sequence. The accuracy rate of rank-2 is 100%, because c1 is in the first two positions of the result sequence. The accuracy rate of rank-5 is 100% because c1 is in the first five positions of the result sequence. Similarly, if the identification results are c2, c1, c3, c4, c5..., the accuracy rate of rank-1 is 0%. Rank −2 has 100% accuracy; The accuracy rate of rank-5 is also 100%. When there are multiple vehicles to be inquired, the average value is the value of rank-m.

2) CMC CURVE

Cumulative Match Characteristic curve is a classical evaluation index in the problem of vehicle re-identification, which is a measure of the performance of the system's ranking ability from 1 to m. The horizontal coordinate of this curve is rank, and the vertical coordinate is the percentage of recognition rate, and it is formed by calculating the hit rate of rank-k, an example like Fig. 25.

3) MAP

Average precision (AP) measures how well the model judges the results on a single query image, while Mean average precision(mAP) measures how well the model judges the results on all query images. mAP is the average of all the AP, AP and mAP can be calculated as follows [75].

$$AP = \frac{\sum_{k=1}^{n} p(k)g(k)}{Ng} \tag{4}$$

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{5}$$

where n is the number of test images and Ng is the number of ground truth images, $p(k)$ is the precision at the k-th position. $g(k)$ represents the indicator function where the value is 1 if match is found at k-th else 0. The mean average precision (mAP) is calculated as follows where Q is number of images queried.

To evaluate the performance of vehicle re-identification methods, this paper classifies the accuracy of these methods on veri-776 and VehicleID datasets, because these two datasets are composed of vehicles with multiple views, Moreover, many vehicle re-identification methods reflect their performance through mAP and Rank values on these two datasets. The accuracy of vehicle re-identification algorithm on veri-776 datasets is shown in Table 5. The vehicle re-identification algorithm is accurate on the VehicleID dataset, as shown in Table 6. The highest values are shown in bold.

IV. CHALLENGES AND POSSIBLE RESEARCH DIRECTIONS

A. CHALLENGES

Although the research work on vehicle re-identification has been carried out for many years, due to limitations by the scale of vehicle datasets and diverse monitoring of the shooting environment, the research work is still challenged. When vehicle re-identification is carried out in traffic surveillance video scene, the resolution of images is different with different cameras, besides, cameras are mounted at different angles, and lighting conditions are different, resulting in difference in angle, scale, and color of the same vehicle in different cameras, which bring great difficulty for vehicle re-identification. Summarize the difficulties and challenges faced in vehicle re-identification work, including:

1) **Limited number of public datasets:** Due to factors such as privacy of vehicle and driver and social security, the scale of publicly available vehicle datasets is not big enough. and the number of vehicles in the same dataset, vehicle type, color, and other attributes are relatively simple.
2) **Small inter-class similarity and large intra-class difference:** Small inter-class similarity is reflected in the similar appearance of vehicles produced by the same automobile brand or different manufacturers. Large intra-class difference is reflected in the fact that the same car looks different due to camera angles, sunshine in the day, lights at night and other factors.
3) **Perspective difference:** Due to the different positions of the cameras in the traffic video monitoring system, the height and angle of the cameras are different, resulting in different perspectives of the same vehicle in different cameras' video frames.
4) **Influence of sunshine in the day, lights at night:** Due to the different lighting conditions of the cameras in the traffic video monitoring system within a day, the color features of the same vehicle photographed by different cameras vary greatly due to the change of lighting conditions. Besides, the vehicle pictures captured at

TABLE 5. The accuracy of vehicle reidentification algorithm on veri-776 dataset.

Method && Reference	Year	mAP (%)	Rank-1(%)	Rank-5 (%)
FACT [131]	2016	19.92	59.65	75.27
FACT + Plate-SNN + STR [54]	2016	27.77	61.44	78.78
XVGAN [94]	2017	24.65	60.20	77.03
Multi-modal metric learning [122]	2017	33.78	60.19	77.40
OIFE+ST [36]	2017	51.42	-	-
Siamese-CNN-Path-LSTM [55]	2017	58.27	83.49	90.04
VGG+classification-oriented loss+triplet loss [72]	2017	58.78	86.41	92.91
ABLN-Ft-16 [113]	2018	24.92	60.49	77.33
SCNN+Ft+CLBL-8-Ft [114]	2018	25.12	60.83	78.55
MSVR [136]	2018	49.30	88.56	-
Space-Time Prior [50]	2018	53.35	82.06	92.31
PROVID [58]	2018	53.42	81.56	95.11
SDC-CNN [59]	2018	53.45	83.49	92.55
JFSDL [57]	2018	53.53	82.90	91.60
RNN-HA [111]	2018	56.80	74.79	87.31
GS-TRE loss W/ mean VGGM [70]	2018	59.47	96.24	98.97
Appearance+Color+Model+Re-Ranking [52]	2018	61.11	89.27	94.76
VAMI+STR [89]	2018	61.32	85.92	91.84
RAM [37]	2018	61.50	88.60	94.00
GAN+LSRO+ re-ranking [91]	2018	64.78	88.62	94.52
SCAN [110]	2019	49.87	82.24	90.76
FDA-Net [92]	2019	55.49	84.27	92.43
Hard-View-EALN [93]	2019	57.44	84.39	94.05
Mob.VFL [49]	2019	58.08	87.18	94.63
DF-CVTC [45]	2019	61.06	91.36	95.77
AAVER [108]	2019	61.18	88.97	94.70
Proposed QD-DLF [118]	2019	61.83	88.50	94.46
SLSR [117]	2019	65.13	91.24	-
VANet [74]	2019	66.34	89.78	95.99
Batch sample [71]	2019	67.55	90.23	96.42
MTML-OSG [125]	2019	68.30	92.00	94.20
MRM [39]	2019	68.55	91.77	95.82
JDRN + re-ranking [127]	2019	73.10	-	-
Part-regularized Near-duplicate [38]	2019	74.30	94.30	98.70
MRL+Softmax Loss [116]	2019	78.50	94.30	**99.00**
PGAN [112]	2019	79.30	96.50	98.30
SSL+re-ranking [90]	2019	89.69	95.41	69.90
PRN+RR [40]	2019	**90.48**	**97.38**	98.87

night are quite different from those taken by the day because of street lamps and other lights at night.

5) **Obscuration:** Due to the uncontrollability of the vehicle's driving route and the diversity of road conditions, the vehicle image under the traffic surveillance camera often has such conditions as road banner obscuration, wire obscuration, and tree branch obstruction.
6) **Scale change:** Because traffic monitoring camera shoot vehicles at different heights or distances, the scale of the vehicle under different monitoring cameras

TABLE 6. The accuracy of vehicle reidentification algorithm on VehicleID dataset.

Method && Reference	Year	Rank-1			Rank-5		
		Small (%)	Medium (%)	Large (%)	Small (%)	Medium (%)	Large (%)
Mixed Diff+CCL [69]	2016	49.00	42.80	38.20	73.50	66.80	61.60
XVGAN [94]	2017	52.87	49.55	44.89	80.83	71.39	66.65
VGG+C+T+S [72]	2017	69.90	66.20	63.20	87.30	82.30	79.40
DJDL [73]	2017	72.30	70.80	68.00	85.70	81.80	78.90
SDC-CNN [59]	2018	56.98	50.57	42.92	86.90	80.05	73.44
NuFACT [58]	2018	48.90	43.64	38.63	69.51	65.34	60.72
JFSDL [57]	2018	54.80	48.29	41.29	85.26	78.79	70.63
RAM [37]	2018	75.20	72.30	67.70	91.50	87.00	84.50
VAMI [89]	2018	63.12	52.87	47.34	83.25	75.12	70.29
A new distance loss [51]	2018	77.10	72.70	70.00	92.80	89.20	87.10
GAN+LSRO+re-ranking [91]	2018	86.50	83.44	81.25	87.38	86.88	84.63
RNN-HA (ResNet+672) [111]	2018	83.80	81.90	81.10	88.10	87.00	87.40
FDA-Net [92]	2019	-	59.84	55.53	-	77.09	74.65
VTGAN [95]	2019	49.48	45.18	40.71	68.66	63.99	59.02
Proposed QD-DLF [118]	2019	72.32	70.66	64.14	92.48	88.90	83.37
Mob.VFL [49]	2019	73.37	69.52	67.41	85.52	81.00	78.48
AAVER [108]	2019	74.69	68.62	63.54	93.82	89.95	85.64
SLSR [117]	2019	75.10	71.80	68.70	89.70	86.10	83.10
DF-CVTC [45]	2019	75.23	72.15	70.46	88.11	84.37	82.13
XG-6-sub-multi [41]	2019	76.10	73.10	71.20	91.20	87.50	84.70
MRM [39]	2019	76.64	74.20	70.86	92.34	88.54	84.82
Part-regularized Near-duplicate [38]	2019	78.40	75.00	74.20	92.30	88.30	86.40
Batch sample [71]	2019	78.80	73.41	69.33	96.17	92.57	89.45
PRN (Single Height-channel Branch) [40]	2019	78.92	74.94	71.58	94.81	92.02	88.46
MRL+Softmax Loss [116]	2019	84.80	80.90	78.40	96.90	94.10	92.10
VANet [74]	2019	88.12	83.17	80.35	**97.29**	**95.14**	**92.97**
GAN+LSRO+re-ranking [90]	2019	**88.67**	**88.31**	**86.67**	91.92	91.81	90.83

is different, which will bring problems like the vehicle is too big to be photographed completely and the vehicle is too small, to cause difficulty in identification.

7) **Resolutions variation:** The standard of traffic surveillance cameras and other factors cause the resolution of the same car to vary greatly, Older cameras tend to have lower resolution and newer ones have higher resolution. Images from earlier cameras have a lower resolution, making it difficult for vehicle re-identification.
8) **Deformation:** The vehicle is deformed by the traffic accident or the shape of the vehicle changes a lot due to different loads.
9) **Background interference:** When the background color of the picture is close to the color of the vehicle, the vehicle re-identification will be disturbed.

B. POSSIBLE RESEARCH DIRECTIONS

As one of the core technologies of intelligent transportation and monitoring, vehicle re-identification technologies play a key role in maintaining social public safety and building smart cities. In recent years, with the in-depth development of deep learning, vehicle re-identification methods based on deep learning have received more and more attention. Combined with existing methods and open vehicle datasets, we present several future research directions from personal opinions based on our survey above, including:

1) **Assistance of spatiotemporal information**: Most of the existing vehicle re-identification methods do not consider the assistance of spatiotemporal information. The search range is from the near time to the far time on the time scale, the search scope extends from the nearby camera to the distant camera on the spatial scale, using the near to far principle to deal with the search process can provide a great help for the vehicle re-identification task. However, in real monitoring scenario, traffic conditions, road map and the weather will affect the vehicle driving route, how to effectively use spatiotemporal clue is still challenging.
2) **Datasets with more information:** The existing vehicle re-identification datasets (VeRi-776 from Beijing University of Posts and Telecommunications, VehicleID in Peking University, and PKU-VD from Peking University) do not provide original video and camera correction information, so they cannot be used for video-based cross-camera vehicle tracking, lacking the ability to track vehicles over a wide area. With the introduction of the CityFlow dataset, it is possible to provide a large-scale tracking of vehicles for future multi-target vehicle tracking. Besides, many datasets do not provide spatio-temporal information, which limits the implementation of the auxiliary vehicle re-identification method by using spatio-temporal information.
3) **Multi-view re-identification:** Compared with person re-identification which is a hot research spot recently, vehicle re-identification mainly faces two major challenges: one is the high variability within the class (Because change of vehicle image caused by change of perspectives is larger than that of person), and the other is the high similarity between the classes. (Because vehicle's appearance produced by different car manufacturers are very similar). In future research work, to solve the first type of challenges and improve the accuracy of vehicle re-identification in different perspectives, firstly, we need to organize large-scale vehicle datasets with multiple perspectives. On this basis, we can consider using the GANs method to learn the correlation between input perspectives and other hidden perspectives through modeling, and learn the transformation model to infer features from other perspectives, that is, a given single-view feature can synthesize multi-view features, and combines metric learning to embed the synthesized multi-view features into the distance space, thereby improving the problem of low vehicle re-identification accuracy under multiple viewing perspectives.
4) **Combine the detection task with the re-identification task:** Current vehicle re-identification missions are based on cropped images of vehicles, in other words, the assumption of the re-identification task is that the boundary box detected by the vehicle is accurate, However, existing testing tasks do not ensure that complete testing is correct, the quality of vehicle detection is likely to affect the accuracy of vehicle re-identification tasks. Therefore, in the future, the detection task and re-identification task can be combined and integrated into an end-to-end framework to analyze and solve the impact of the deviation of detection task on the accuracy of the vehicle re-identification task, it is still a challenge.
5) **Integrate multiple approaches to improve the accuracy of vehicle re-identification:** Many researchers consider vehicle re-identification as a classification task for fine-grained vehicle identification, but it cannot be accurately matched. Considering the characteristics of classification learning and metric learning, classification task training is easy, and metric learning is good at detail subdivision. Combining the characteristics of different methods and effectively integrate them can realize the complementary advantages of the methods and improve the accuracy of vehicle re-identification.
6) **Effectively applied the vehicle re-identification technology to real traffic scenes by the assistant of methods such as transfer learning:** Features learned from one dataset may not apply to another due to differences in data distribution, trained models on existing vehicle datasets may not be fully applicable to real traffic scenarios. Besides, the types of cars in different cities may be different, it is still a challenge to apply the vehicle re-identification model to real traffic scenes through the training of existing datasets, and further research is needed. Transfer Learning allows the domain of training to be different from the domain data distribution of testing, by using methods such as transfer learning, the vehicle re-identification technology can be applied to real traffic scenes effectively.

V. CONCLUSION

With the improvement of social public infrastructure, the number of vehicles on the roads has increased year by year, which has led to higher requirements for the ability to analyze vehicles shot at surveillance cameras. The emergence of vehicle re-identification technology meets the demand for public safety and construction of smart cities, as well as provide guarantees for the comprehensive improvement of traffic management and service levels. Vehicle re-identification has been widely concerned in recent years. In this paper, we focus on vehicle re-identification methods based on deep learning, and categorize these methods into five categories, i.e. methods based on local features, methods based on representation learning, methods based on metric learning, methods based on unsupervised learning, and methods based on attention mechanism. Furthermore, we compare these methods from characteristics, advantages, and disadvantages. The vehicle public datasets have a great influence on the accuracy of vehicle re-identification, therefore, this paper summarizes the vehicle datasets for vehicle re-identification in recent years, and give a list of the scale, number of vehicles, number

of models, number of colors, characteristics, and application for each dataset. Besides, some common strategies used to evaluate the performance of the vehicle re-identification algorithm are introduced, and we compare some vehicle re-identification method's accuracy in veri-776 and VehicleID datasets. At last, we summarize the difficulties and challenges faced by vehicle re-identification and discuss possible research directions in the future. Through the survey of vehicle re-identification methods based on deep learning, hoping to provide a guideline and assistances for future research.

REFERENCES

[1] A. Cocchia, "Smart and digital city: A systematic literature review," in *Smart City: How to Create Public and Economic Value with High Technology in Urban Space*. Switzerland: Springer, 2014, pp. 13–43, doi: 10.1007/978-3-319-06160-3_2.
[2] B. F. Momin and T. M. Mujawar, "Vehicle detection and attribute based search of vehicles in video surveillance system," in *Proc. ICCPCT*, Nagercoil, India, Mar. 2015, pp. 1–4.
[3] W. Chu, Y. Liu, C. Shen, D. Cai, and X.-S. Hua, "Multi-task vehicle detection with region-of-interest voting," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 432–441, Jan. 2018.
[4] Z. He, Y. Lei, S. Bai, and W. Wu, "Multi-Camera vehicle tracking with powerful visual features and spatial-temporal cue," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 203–212.
[5] P. Li, G. Li, Z. Yan, Y. Li, M. Lu, P. Xu, Y. Gu, B. Bai, and Y. Zhang, "Spatio-temporal consistency and hierarchical matching for multi-target multi-camera vehicle tracking," in *Proc. CVPR Workshops*, Long Beach, CA, USA, May 2019, pp. 222–230.
[6] J. Špaňhel, J. Sochor, and A. Makarov, "Vehicle fine-grained recognition based on convolutional neural networks for real-world applications," in *Proc. 14th Symp. NEUREL*, Belgrade, Serbia, Nov. 2018, pp. 1–5.
[7] Z. Chen, C. Ying, C. Lin, S. Liu, and W. Li, "Multi-view vehicle type recognition with feedback-enhancement multi-branch CNNs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2590–2599, Sep. 2019.
[8] B. Hu, J.-H. Lai, and C.-C. Guo, "Location-aware fine-grained vehicle type recognition using multi-task deep networks," *Neurocomputing*, vol. 243, no. 21, pp. 60–68, Jun. 2017.
[9] 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.
[10] J. Zakria, J. Cai, J. Deng, M. U. Aftab, M. S. Khokhar, and R. Kumar, "Efficient and deep vehicle re-identification using multi-level feature extraction," *Appl. Sci.*, vol. 9, no. 7, p. 1291, 2019.
[11] W.-H. Lin and D. Tong, "Vehicle re-identification with dynamic time windows for vehicle passage time estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1057–1063, Dec. 2011.
[12] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya, "Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors," *Transp. Res. C, Emerg. Technol.*, vol. 17, no. 6, pp. 586–606, Dec. 2009.
[13] S. M. Silva and C. r. Jung, "License plate detection and recognition in unconstrained scenarios," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 593–609.
[14] N. Watcharapinchai and S. Rujikietgumjorn, "Approximate license plate string matching for vehicle re-identification," in *Proc. AVSS*, Lecce, Italy, Aug./Sep. 2017, pp. 1–6.
[15] R. S. Feris, B. Siddiquie, J. Petterson, Y. Zhai, A. Datta, L. M. Brown, and S. Pankanti, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 28–42, Feb. 2012.
[16] B. C. Matei, H. S. Sawhney, and S. Samarasekera, "Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features," in *Proc. CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 3465–3472.
[17] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *Comput. Vis. Image Understand.*, vol. 182, pp. 50–63, May 2019.
[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, Kerkyra, Greece, 1999, pp. 1150–1157.
[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, San Diego, CA, USA, Jun. 2005, pp. 886–893.
[20] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[21] D. Zapletal and A. Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Proc. CVPR Workshops*, Las Vegas, NV, USA, Jun. 2016, pp. 1568–1574.
[22] H. C. Chen, J.-W. Hsieh, and S.-P. Huang, "Real-Time vehicle re-identification system using symmelets and HOMs," in *Proc. 15th AVSS*, Auckland, New Zealand, Nov. 2018, pp. 1–6.
[23] M. Cormier, L. W. Sommer, and M. Teutsch, "Low resolution vehicle re-identification based on appearance features for wide area motion imagery," in *Proc. WACVW*, Lake Placid, NY, USA, Mar. 2016, pp. 1–7.
[24] S. Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in *Proc. NIPS*, Vancouver, BC, Canada, Dec. 2001, pp. 831–837.
[25] I. G. Iyyakutti and S. Prakash, "False mapped feature removal in spin images based 3D ear recognition," in *Proc. SPIN*, Noida, India, Feb. 2016, pp. 620–623.
[26] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, Berlin, Germany, 2006, pp. 404–417.
[27] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, Sep. 2005.
[28] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. CVPR*, Miami, FL, USA, Jun. 2009, pp. 1932–1939.
[29] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[33] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 1666–1674.
[34] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," Mar. 2017, *arXiv:1603.06765*. [Online]. Available: https://arxiv.org/abs/1603.06765
[35] H. Deng, T. Birdal, and S. Ilic, "PPFNet: Global context aware local features for robust 3D point matching," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 195–205.
[36] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. ICCV*, Venice, Italy, Jun. 2017, pp. 379–387.
[37] X. Liu, S. Zhang, Q. Huang, and W. Gao, "RAM: A region-aware deep model for vehicle re-identification," in *Proc. ICME*, San Diego, CA, USA, Jul. 2018, pp. 1–6.
[38] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 3997–4005.
[39] J. Peng, H. Wang, T. Zhao, and X. Fu, "Learning multi-region features for vehicle re-identification with context-based ranking method," *Neurocomputing*, vol. 259, pp. 427–437, Sep. 2019.
[40] H. Chen, B. Lagadec, and F. Bremond, "Partition and reunion: A two-branch neural network for vehicle re-identification," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 184–192.
[41] Y. Zhao, C. Shen, H. Wang, and S. Chen, "Structural analysis of attributes for vehicle re-identification and retrieval," *IEEE Trans. Intell. Transp. Syst.*, to be published.
[42] X. Ma, K. Zhu, H. Guo, J. Wang, M. Huang, and Q. Miao, "Vehicle re-identification with refined part model," in *Proc. ICMEW*, Shanghai, China, Jul. 2019, pp. 603–606.

[43] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[44] Y. Li, L. Zhuo, X. Hu, and J. Zhang, "A combined feature representation of deep feature and hand-crafted features for person re-identification," in *Proc. PIC*, Shanghai, China, Dec. 2016, pp. 224–227.

[45] A. Zheng, X. Lin, C. Li, R. He, and J. Tang, "Attributes guided feature learning for vehicle re-identification," Mar. 2019, *arXiv:1905.08997*. [Online]. Available: https://arxiv.org/abs/1905.08997

[46] P. Huang, R. Huang, J. Huang, R. Yangchen, Z. He, X. Li, and J. Chen, "Deep feature fusion with multiple granularity for vehicle re-identification," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 80–88.

[47] J.-H. Hou, H.-Q. Zeng, L. Cai, J.-Q. Zhu, and J. Chen, "Random occlusion assisted deep representation learning for vehicle re-identification," *Control Theory Appl.*, vol. 35, no. 12, pp. 1725–1730, Dec. 2018.

[48] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. ICCV Workshops*, Sydney, NSW, Australia, Jun. 2013, pp. 554–561.

[49] S. A. S. Alfasly, "Variational representation learning for vehicle re-identification," in *Proc. ICIP*, Taipei, China, 2019, pp. 3118–3122.

[50] C.-W. Wu, C.-T. Liu, C.-E. Chiang, W.-C. Tu, and S.-Y. Chien, "Vehicle re-identification with the space-time prior," in *Proc. CVPR Workshops*, Salt Lake City, UT, USA, Jun. 2018, pp. 121–128.

[51] L. Wei, X. Liu, J. Li, and S. Zhang, "VP-ReID: Vehicle and person re-identification system," in *Proc.8th ACM ICMR*, Yokohama, Japan, Jun. 2018, pp. 501–504.

[52] N. Jiang, Y. Xu, Z. Zhou, and W. Wu, "Multi-Attribute driven vehicle re-identification with spatial-temporal re-ranking," in *Proc. ICIP*, Athens, Greece, Oct. 2018, pp. 858–862.

[53] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Proc. NIPS*, Vancouver, BC, Canada, Dec. 2002, pp. 521–528.

[54] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. ECCV*, Amsterdam, The Netherlands, Oct. 2016, pp. 869–884.

[55] 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. ICCV*, Venice, Italy, Jun. 2017, pp. 1918–1927.

[56] C. Cui, N. Sang, C. Gao, and L. Zou, "Vehicle re-identification by fusing multiple deep neural networks," in *Proc. IPTA*, Montreal, QC, Canada, Nov./Dec. 2017, pp. 1–6.

[57] J. Zhu, H. Zeng, Y. Du, Z. Lei, L. Zheng, and C. Cai, "Joint feature and similarity deep learning for vehicle re-identification," *IEEE Access*, vol. 6, pp. 43724–43731, 2018.

[58] 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.

[59] J. Zhu, H. Zeng, Z. Lei, S. Liao, L. Zheng, and C. Cai, "A shortly and densely connected convolutional neural network for vehicle re-identification," in *Proc. 24th ICPR*, Beijing, China, Aug. 2018, pp. 3285–3290.

[60] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. 24th CVPR*, Boston, MA, USA, Jun. 2015, pp. 815–823.

[61] 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.

[62] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. CVPR*, Columbus, OH, USA, Jun. 2014, pp. 1386–1393.

[63] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 1320–1329.

[64] 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, Apr. 2017, Art. no. 27.

[65] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 1114–1123.

[66] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, Amsterdam, The Netherland, 2016, pp. 499–515.

[67] 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.

[68] 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.

[69] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 2167–2175.

[70] 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.

[71] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: An efficient baseline using triplet embedding," Aug. 2019, *arXiv:1901.01015*. [Online]. Available: https://arxiv.org/abs/1901.01015

[72] Y. Zhang, D. Liu, and Z.-J. Zha, "Improving triplet-wise training of convolutional neural network for vehicle re-identification," in *Proc. ICME*, Hong Kong, Jul. 2017, pp. 1386–1391.

[73] Y. Li, Y. Li, H. Yan, and J. Liu, "Deep joint discriminative learning for vehicle re-identification and retrieval," in *Proc. ICIP*, Beijing, China, Sep. 2017, pp. 395–399.

[74] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. ICCV*, Seoul, South Korea, Jun. 2019, pp. 8282–8291.

[75] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, "VR-PROUD: Vehicle re-identification using progressive unsupervised deep architecture," *Pattern Recognit.*, vol. 90, pp. 52–65, Jun. 2019.

[76] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, Portland, OR, USA, Jun. 2013, pp. 3586–3593.

[77] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 994–1003.

[78] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 2275–2284.

[79] P. A. Marín-Reyes, A. Palazzi, L. Bergamini, S. Calderara, J. Lorenzo-Navarro, and R. Cucchiara, "Unsupervised vehicle re-identification using triplet networks," in *Proc. CVPR Workshops*, Salt Lake City, UT, USA, Jun. 2018, pp. 166–171.

[80] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. ICCV*, Santiago, Chile, Jun. 2015, pp. 2794–2802.

[81] R. M. S. Bashir, M. Shahzad, and M. M. Fraz, "DUPL-VR: Deep unsupervised progressive learning for vehicle re-identification," in *Proc. ISVC*, Las Vegas, NV, USA, 2018, pp. 286–295.

[82] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, QC, Canada, 2014, pp. 2672–2680.

[83] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," Oct. 2017, *arXiv:1710.10196*. [Online]. Available: https://arxiv.org/abs/1710.10196

[84] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Nov. 2015, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434

[85] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 1125–1134.

[86] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 2172–2180.

[87] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. ICML*, Sydney, NSW, Australia, 2017, pp. 2642–2651.

[88] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 2242–2251.

[89] Y. Zhou and L. Shao, "Viewpoint-Aware attentive multi-view inference for vehicle re-identification," in *Proc. CVPR*, Salt Lake City, UT, USA, Jun. 2018, pp. 6489–6498.

[90] F. Wu, S. Yan, J. S. Smith, and B. Zhang, "Vehicle re-identification in still images: Application of semi-supervised learning and re-ranking," *Signal Process., Image Commun.*, vol. 76, pp. 261–271, Aug. 2019.
[91] F. Wu, S. Yan, J. S. Smith, and B. Zhang, "Joint semi-supervised learning and re-ranking for vehicle re-identification," in *Proc. 24th ICPR*, Beijing, China, Aug. 2018, pp. 278–283.
[92] 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. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 3235–3243.
[93] 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.
[94] Y. Zhou and L. Shao, "Cross-view GAN based vehicle generation for re-identification," in *Proc. 28th BMVC*, London, U.K., 2017, pp. 1–12.
[95] J. Peng, H. Wang, T. Zhao, and X. Fu, "Cross domain knowledge transfer for unsupervised vehicle re-identification," in *Proc. ICMEW*, Shanghai, China, 2019, pp. 453–458.
[96] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, Montreal, QC, Canada, 2015, pp. 2017–2025.
[97] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2913372.
[98] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.
[99] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. NIPS*, Montreal, QC, Canada, 2014, pp. 2204–2212.
[100] Y. Li and Y. Wang, "A multi-label image classification algorithm based on attention model," in *Proc. ICIS*, Singapore, Jun. 2018, pp. 728–731.
[101] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458.
[102] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 4476–4484.
[103] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 5219–5227.
[104] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 1251–1259.
[105] F. Fang, H. Wang, and P. Tang, "Image captioning with word level attention," in *Proc. ICIP*, Athens, Greece, Oct. 2018, pp. 1278–1282.
[106] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 1300–1309.
[107] M.-C. Chang, J. Wei, Z.-A. Zhu, Y.-M. Chen, C.-S. Hu, M.-X. Jiang, and C.-K. Chiang, "AI city challenge 2019–city-scale video analytics for smart transportation," in *Proc. CVPR Workshops*, Long Beach, CA, USA, May 2019, pp. 99–108.
[108] 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. ICCV*, Seoul, South Korea, Jun. 2019, pp. 6132–6141.
[109] P. Khorramshahi, N. Peri, A. Kumar, A. Shah, and R. Chellappa, "Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 239–246.
[110] S. Teng, X. Liu, S. Zhang, and Q. Huang, "SCAN: Spatial and channel attention network for vehicle re-identification," in *Proc. Pacific Rim Conf. Multimedia*, vol. 11166, 2018, pp. 350–361.
[111] X.-S. Wei, C.-L. Zhang, L. Liu, C. Shen, and J. Wu, "Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification," in *Proc. ACCV*, Perth, WA, Australia, Dec. 2018, pp. 575–591.
[112] X. Zhang, R. Zhang, J. Cao, D. Gong, M. You, and C. Shen, "Part-guided attention learning for vehicle re-identification," Sep. 2019, *arXiv:1909.06023*. [Online]. Available: https://arxiv.org/abs/1909.06023
[113] Y. Zhou and L. Shao, "Vehicle re-identification by adversarial bi-directional LSTM network," in *Proc. WACV*, Lake Tahoe, NV, USA, Mar. 2018, pp. 653–662.
[114] 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. 3275–3287, Jul. 2018.
[115] F. Zheng, X. Miao, and H. Huang, "Fast vehicle identification via ranked semantic sampling based embedding," in *Proc. 27th IJCAI*, Stockholm, Sweden, Jul. 2018, pp. 3697–3703.
[116] W. Lin, Y. Li, X. Yang, P. Peng, and J. Xing, "Multi-View learning for vehicle re-identification," in *Proc. ICME*, Shanghai, China, Jul. 2019, pp. 832–837.
[117] Y. Xu, N. Jiang, L. Zhang, Z. Zhou, and W. Wu, "Multi-scale vehicle re-identification using self-adapting label smoothing regularization," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 2117–2121.
[118] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Vehicle re-identification using quadruple directional deep learning features," *IEEE Trans. Intell. Transp. Syst.*, to be published.
[119] J. Zhu, H. Zeng, X. Jin, Y. Du, L. Zheng, and C. Cai, "Joint horizontal and vertical deep learning feature for vehicle re-identification," *Sci. China Inf. Sci.*, vol. 62, no. 9, pp. 199101:1–199101:2, Sep. 2019.
[120] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. ICCV*, Seoul, South Korea, Jun. 2019, pp. 211–220.
[121] D. Liang, K. Yan, Y. Wang, W. Zeng, Q. Yuan, X. Bao, and Y. Tian, "Deep hashing with multi-task learning for large-scale instance-level vehicle search," in *Proc. ICMEW*, Hong Kong, Jul. 2017, pp. 192–197.
[122] Y. Tang, D. Wu, Z. Jin, W. Zou, and X. Li, "Multi-modal metric learning for vehicle re-identification in traffic surveillance environment," in *Proc. ICIP*, Beijing, China, Sep. 2017, pp. 2254–2258.
[123] J. Hou, H. Zeng, J. Zhu, J. Hou, J. Chen, and K.-K. Ma, "Deep quadruplet appearance learning for vehicle re-identification," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8512–8522, Sep. 2019.
[124] T.-W. Huang, J. Cai, H. Yang, H.-M. Hsu, and J.-N. Hwang, "Multi-view vehicle re-identification using temporal attention model and metadata re-ranking," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 434–442.
[125] A. Kanaci, M. Li, S. Gong, and G. Rajamanoharan, "Multi-task mutual learning for vehicle re-identification," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 62–70.
[126] A. Kanacı, X. Zhu, and S. Gong, "Vehicle re-identification by fine-grained cross-level deep learning," in *Proc. AMMDS Workshop*, London, U.K., 2017, pp. 1–6.
[127] C.-T. Liu, M.-Y. Lee, C.-W. Wu, B.-Y. Chen, T.-S. Chen, Y.-T. Hsu, and S.-Y. Chien, "Supervised joint domain learning for vehicle re-Identification," in *Proc. CVPR Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 45–52.
[128] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 3006–3015.
[129] J. Sochor, J. Špaňhel, and A. Herout, "BoxCars: Improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 97–108, Jan. 2019.
[130] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proc. ICCV*, Venice, Italy, Jun. 2017, pp. 562–570.
[131] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. ICME*, Seattle, WA, USA, Jul. 2016, pp. 1–6.
[132] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. CVPR*, Boston, MA, USA, Jun. 2015, pp. 3973–3981.
[133] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 6853–6860.
[134] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," Mar. 2019, *arXiv:1903.09254*. [Online]. Available: https://arxiv.org/abs/1903.09254

[135] X. Li, M. Yuan, Q. Jiang, and G. Li, "VRID-1: A basic vehicle re-identification dataset for similar vehicles," in *Proc. ITSC*, Yokohama, Japan, Oct. 2017, pp. 1–8.
[136] A. Kanacı, X. Zhu, and S. Gong, "Vehicle re-identification in context," in *Proc. GCPR*, Stuttgart, Germany, 2018, pp. 377–390.
[137] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *Proc. ICCV*, Seoul, South Korea, 2019, pp. 460–469.

HONGBO WANG received the B.S. degree in computer software from Hebei University, China, in 1998, and the Ph.D. degree in computer application technology from the Beijing University of Posts and Telecommunications (BUPT), China, in 2006. He is currently an Associate Professor with the State Key Laboratory of Networking and Switching Technology, BUPT. His main research interests include computer vision, cloud computing, big data, data center network, and the Internet measurement, in which he has published over 60 technical articles in referred journals and conference proceedings.

JIAYING HOU received the B.S. degree in software engineering from Hebei University, China, in 2018. She is currently pursuing the degree with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China. Her research interests include computer vision and deep learning.

NA CHEN received the M.S. degree in computer science and technology from the Beijing University of Posts and Telecommunications, in 2019. From 2017 to 2019, she studied the field of vehicle re-identification. She is currently a Designer with China Aerospace Science and Industry Corporation. Her research interests include artificial intelligence and cloud computing.