Chapter

Object Re-Identification Based on Deep Learning

Xiying Li and Zhihao Zhou

Abstract

With the explosive growth of video data and the rapid development of computer vision technology, more and more relevant technologies are applied in our real life, one of which is object re-identification (Re-ID) technology. Object Re-ID is currently concentrated in the field of person Re-ID and vehicle Re-ID, which is mainly used to realize the cross-vision tracking of person/vehicle and trajectory prediction. This chapter combines theory and practice to explain why the deep network can re-identify the object. To introduce the main technical route of object Re-ID, the examples of person/vehicle Re-ID are given, and the improvement points of existing object Re-ID research are described separately.

Keywords: object re-identification, deep learning, person re-identification, vehicle re-identification, feature extraction

1. Introduction

In a surveillance camera without overlapping vision, a recognized object is identified again after imaging conditions (including monitoring scene, lighting conditions, object pose, etc.) change, which is called object re-identification (Object Re-ID). Object Re-ID technology has important research significance in intelligent monitoring, multi-object tracking and other fields. In recent years, scholars have paid extensive attention to it. The main application areas of object Re-ID are person Re-ID and vehicle Re-ID.

Person re-identification (Re-ID) is a technology that uses computer vision technology to judge whether there is a specific person in the image or video sequence. It is widely regarded as a sub-problem of image retrieval. Given a monitor person image, retrieve the image of the row of people across the device. It aims to make up for the visual limitations of the current fixed cameras, and can be combined with person detection and pedestrian tracking technology, which can be widely used in intelligent video monitoring, intelligent security and other fields.

Vehicle re-identification (Re-ID) aims to quickly search, locate and track the target vehicles across surveillance camera networks, which plays key roles in maintaining social public security and serves as a core module in the large-scale vehicle recognition, intelligent transportation, surveillance video analytic platforms. Vehicle Re-ID refers to the problem of identifying the same vehicle in a large scale vehicle database given a probe vehicle image. In particular, vehicle Re-ID can be regarded as a fine-grained recognition task that aims at recognizing the subordinate category of a given class. The wide popularization and use of road video

monitoring makes vehicle matching based on video image become the hot spot in current intelligent traffic research, and the typical applications are vehicle origin-destination analysis and vehicle trajectory reconstruction. In some cases which license plate number could be recognized clearly and accurately, vehicle Re-ID could be realized by match the license plate number. However, in more cases, such as license plate can't be recognized (for most surveillance video), license plate occlusion and so on in the criminal investigation, it is necessary to realize the vehicle Re-ID without license plates by using computer vision and other related technologies.

2. Related work of object Re-ID

As an emerging research topic, object Re-ID has attracted great efforts. Existing research directions of object Re-ID are mainly divided into person Re-ID and vehicle Re-ID. In this section, we will review the relevant works from person Re-ID and vehicle Re-ID.

2.1 Person Re-ID

We will review the relevant work [1] of person Re-ID from following aspects: person Re-ID based on representation learning, metric learning, local features and video sequence.

2.1.1 Person Re-ID based on representation learning

Methods based on representation learning are a kind of very common person Re-ID methods, which is mainly thanks to the deep learning, especially the Convolutional neural network (CNN) development. Sunderrajan et al. [2] propose a clothing context-aware color extraction method to learn color drift patterns in a non-parametric manner using the random forest distance (RFD) function. Geng et al. [3] proposed a person Re-ID algorithm which used Classification loss and verification loss to train the network (including Classification Subnet and Verification Subnet), and the network inputs several pairs of pedestrian images. The classification subnetwork makes ID prediction on the image, and calculates the classification error loss according to the predicted ID. The sub-network integrates the features of two images and judge whether these two images belong to the same pedestrian. The sub-network is essentially equivalent to a binary classification network. After enough data training, input a test image again, and the network will automatically extract a feature, which is used for person Re-ID. For the problem that pedestrian ID information alone is not enough to learn a model with strong generalization ability, the researchers added attributes such as gender, hair and clothing to the pedestrian images. By introducing the pedestrian attribute label, the model should not only accurately predict the pedestrian ID, but also predict the correct pedestrian attributes, which greatly increases the generalization ability of the model. Most papers also show that this method is effective. Lin et al. [4] proposed a person Re-ID algorithm based on multiple attributes. In this algorithm, the features of network output are not only used to predict the ID information of pedestrians, but also to predict the attributes of each pedestrian. The combination of ID loss and attribute loss can improve the generalization ability of the network. Currently, there is still a lot of work based on representational learning. Representational learning has also become a very important baseline of Re-ID field. Moreover, the method of representational learning is more robust, the training is more

stable, and the results are easier to reproduce. However, representation learning is easy to be overfitted in the domain of the data set, and when the training ID is increased to a certain extent, it will be weak.

2.1.2 Person Re-ID based on metric learning

Metric learning is a method widely used in the field of image retrieval. Unlike representational learning, metric learning aims to learn the similarity between two images through the network. In the problem of person Re-ID, the similarity of different images of the same pedestrian is greater than that of different images of the different pedestrians. Finally, the loss function of the network makes the distance between the same pedestrian images (positive sample pairs) as small as possible, and the distance between different pedestrian images (negative sample pairs) as large as possible. Common measures of learning loss include Contrastive loss, Triplet loss, Quadruplet loss, Triplet hard loss with batch hard mining (TriHard loss) and Margin sample mining loss (MSML). Varior et al. [5] proposed Siamese Network, and trained the network model by contrast loss. By reducing the contrast loss, the distance between positive sample pairs is gradually reduced, and the distance between negative sample pairs is gradually increased, so as to meet the need of person Re-ID. Triplet loss is a widely used metric learning loss and a lot of metric learning methods have evolved based on triples. Ding et al. [6] considered the re-identification problem as a ranking issue and used triplet loss to obtain the relative distance between images. Chen et al. [7] designed a quadruplet loss process, which can lead to model outputs with larger inter-class variation and smaller intra-class variation compared with the triplet loss method. Hermans et al. [8] proposed a batch training based online difficult sample sampling method, which is named TriHard Loss. Traditional triplet sample mining strategy randomly select three images from training data, and most of the sampled images are simple and easily distinguishable sample pairs, which is not conducive to better representation of network learning. This paper proposes a sample mining strategy that can obtain more difficult samples which can improve the generalization ability of the network. Xiao et al. [9] proposed Margin sample mining loss which introduces the idea of hard sample sampling. MSML losses are calculated by picking only the hardest positive sample pair and the hardest negative sample pair. It is a measure learning method that takes into account both relative distance and absolute distance and introduces the idea of difficult sample sampling.

2.1.3 Person Re-ID based on local features

In the early stage of ReID's research, people still focused on global feature, but later the global feature encountered a bottleneck, so they began to study local feature gradually. The commonly used methods to extract local features include image segmentation, positioning of skeleton key points and posture correction, etc. Image segmentation is a very common way to extract local features. Wei et al. [10] develop a pedestrian image descriptor named Global-Local-Alignment Descriptor, this descriptor explicitly leverages the local and global cues in human body to generate a discriminative and robust representation. In order to solve the failure of manual image slice in the case of image misalignment, some papers first align pedestrians with some prior knowledge, which mainly includes pre-trained human Pose and Skeleton key points model. Su et al. [11] proposed a pose-driven deep convolutional model to alleviate the pose variations and learn robust feature representations from both the global images and different local parts. Liang et al. [12] first estimated the key points of pedestrians with the model of attitude estimation,

and then made the same key points align with affine transformation. To extract local features at different scales, they set three different PoseBox combinations; afterwards, the three PoseBox corrected images were sent to the network together with the original corrected images to extract features, which contained both global and local information. In order to solve the problem of local feature alignment, most methods need an additional skeleton key point or pose estimation model. Zhang et al. [13] proposed an automatic alignment model based on SP distance (AlignedReID), which automatically aligned local features without requiring additional information.

2.1.4 Person Re-ID based on video sequence

The main difference between video sequence-based methods is that such methods not only consider the content information of the image, but also consider the motion information between frames. Liu et al. [14] propose an algorithm called Accumulative motion context network (AMOC), the input of AMOC includes the original image sequence and the extracted optical flow sequence. AMOC has Spatial network and Motion network. Each frame of an image sequence is input into Spat Nets to extract the global content features of the image, the two adjacent frames will be sent to the Moti Nets to extract the optical flow pattern features; then the spatial features and optical flow features are merged and input into an RNN to extract the temporal features. Through the AMOC network, each image sequence can be extracted with a feature that integrates content information and motion information. The network adopts classification loss and comparison loss to train the model. Sequential image features with motion information can improve the accuracy of person Re-ID. Mazzeo et al. [15] propose a multi camera architecture for wide area surveillance and a real time people tracking algorithm across non overlapping cameras, they proposed different methodologies [16] to extract the color histogram information from each object patches for the intra-camera and compared different methods to evaluate the colour Brightness Transfer Function (BTF) between non overlapping cameras for inter-camera tracking. This method outperforms the performance in terms of matching rate between different cameras.

2.2 Vehicle Re-ID

We will review the relevant works of vehicle Re-ID from three aspects: vehicle re-identification based on artificial design feature, vehicle re-identification based on deep learning feature and vehicle re-identification based on fusion feature.

2.2.1 Vehicle Re-ID based on artificial design feature

In the initial vehicle matching problem, sensor tag matching is adopted. Tian et al. [17] proposed an algorithm for vehicle Re-ID based on multiple sensor nodes. According to the matching results of the same vehicle label obtained by different nodes, the vehicle state was determined and the label segmentation was modified. Meanwhile, the time difference between vehicles was modified according to the relationship between different acquired labels. Coifman [18] proposed a matching algorithm for individual vehicles measured on the highway detector and made corresponding measurements on another detector upstream. Rios-Cabrera et al. [19] proposed a comprehensive scheme for solving the problems of vehicle detection, recognition and tracking in view of the practical application of tunnel monitoring, and proposed compact binary features to improve the recognition effect for the influence of poor imaging conditions and vehicle lights in tunnel monitoring.

Due to the late rise of vehicle Re-ID research, when traditional methods have not been applied to this problem too much, the deep learning technology has developed in a big bang. Almost all subsequent studies are based on deep learning technology, which greatly improves the effect of Re-ID.

2.2.2 Vehicle Re-ID based on deep learning feature

In recent years, convolutional neural network has been widely used in the field of computer vision and achieved remarkable effects. Because the depth features extracted by deep convolutional networks have stronger description ability, more and more scholars have applied them in vehicle Re-ID. Liu et al. [20] proposed a large-scale vehicle Re-ID data set "VeRi," and puts forward a method of feature Fusion FACT by combining the depth of the vehicle network features, color features and SIFT features to match the same vehicle, the follow-up of vehicle recognition of other study, a large number of experiment based on the data set, thereby evaluating effectiveness and superiority of the proposed algorithm. Liu et al. [21] solved the problem of difficulty in triplet loss convergence by adding a feature representation between the sample and each individual vehicle into the triplet network to model intra-class variance. Li et al. [22] proposed DJDL (Deep Joint Discriminative Learning) model, which projects the original vehicle image into Euclidean space through a two-branch Deep convolution network. Zhang et al. [23] proposed a guided Triplet network, which added classification loss to the original triplet loss function and strongly restricted the original training network, thus improving the Re-ID efficiency. Marin et al. [24] designed a metric learning model based on the supervision of the local constraints, its use in pairs and triple constraints to train a network, the network is able to share the same identity of the sample distribution of high similarity, and keep a distance of different identity in the feature space, the algorithm is one of the biggest advantage is to use the vehicle tracking to automatically generate a set of weak tag data, and will automatically generate data sets used in depth training network to complete the vehicles Re-ID task.

2.2.3 Vehicle Re-ID based on fusion feature

For monitoring video, in addition to appearance information of images, information other than appearance features (such as, space-time information) is also of great mining significance. Liu et al. [25] proposed a segmented vehicle Re-ID algorithm, which first used appearance features for preliminary screening, then used license plate information for matching, and finally used spatial and temporal information for reordering. After the method was integrated with spatial and temporal information, the effect was improved to a certain extent. Jiang et al. [26] proposed a vehicle Re-ID algorithm based on multiple attribute training and sort by spatial-temporal similarity, the vehicle image color, models, vehicle feature extraction with individual respectively. Through the fusion of multiple features for the initial Re-ID, the Re-ID results are reordered by the spatial-temporal similarity, and good results are obtained. Shen et al. [27] proposed a two-stage architecture containing complex spatiotemporal information, given a pair of vehicle images with spatio-temporal information, candidate visual spatio-temporal paths (where each visual spatio-temporal state corresponds to an actual vehicle image with spatio-temporal information) are generated by an MRF chain model with a deep learning function, and then candidate paths and paired queries are used to generate their similarity scores for the model. In addition to fusion of information other than the apparent features of images, many scholars have also studied fusion of manual features and deep convolution features, fusion of various attribute features or feature fusion

between different image regions. Li et al. [28] proposed a vehicle Re-ID algorithm based on fusion features extract from different part of vehicle, firstly, a part detection algorithm [29] is used to obtain the attention area with big difference between different vehicles. Then, feature extraction was carried out on the detected area, and features of the two areas were fused to generate new fusion features. Liang et al. [30] put forward a new method of supervision and the depth of the hash to handle large-scale vehicle search problem, the use of multitasking learning to learn, vehicle model, vehicle image color depth features of individual ID hash code, the experimental results show the effectiveness of the proposed method, the method in classification loss and triple loss case depth hash method is superior to single task.

3. Some public database for object Re-ID

With the development of Re-ID research, many scholars have published the data sets of relevant fields. The following are some commonly used person Re-ID data sets and vehicle Re-ID data sets.

3.1 Person Re-ID data sets

Person Re-ID data sets commonly used in deep learning methods include VIPeR [31], PRID2011 [32], CUHK03 [33], Market1501 [34], CUHK-SYSU [35], MARS [36], DukeMTMC-reID [37]. In addition to the common data sets that are already open source, there are several newer data sets, such as SYSU-MM01 [38], LPW [39], MSMT17 [40], LVreID [41], the download link is not yet open. The following is a detailed description of CUHK03 and Market1501.

3.1.1 CUHK03

The dataset includes 13,164 images of 1360 pedestrians. The whole dataset is captured with six surveillance cameras. Each identity is observed by two disjoint camera views and has an average of 4–8 images in each view. Some examples are shown in **Figure 1**. Besides the scale, it has the following characteristics.

This dataset is partitioned into training set (1160 persons), validation set (100 persons), and test set (100 persons). Each person has roughly 4–8 photos per view, which means there are almost 26,000 positive training pairs before data augmentation.

3.1.2 Market1501

During dataset collection, a total of six cameras were placed in front of a campus supermarket, including five 1280 × 1080 HD cameras, and one 720 × 576 SD camera. Overlapping exists among these cameras. This dataset contains 32,668 boxes of 1501 identities. Due to the open environment, images of each identity are captured by at most six cameras. Each annotated identity is captured by at least two cameras, so that cross-camera search can be performed. Overall, the dataset has the following featured properties.

The dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively. During testing, for each identity, it selects one query image in each camera. Note that, the selected queries are hand-drawn, instead of DPM-detected as in the gallery. The reason is that in reality, it is very convenient to interactively draw a box, which can yield higher recognition accuracy. The search process is performed in a cross-camera mode, i.e., relevant

Figure 1.
Person samples selected from the CUHK03 dataset.

Figure 2.
Person samples selected from the Market1501 dataset.

images captured in the same camera as the query are viewed as "junk." In this scenario, an identity has at most six queries, and there are 3368 query images in total. Dataset examples are shown in **Figure 2**.

3.2 Vehicle Re-ID data sets

Vehicle Re-ID data sets commonly used in deep learning methods include VRID-1 [42], VeRi-776 [25], VehicleID [21].

3.2.1 VRID-1

The open dataset VRID-1 for vehicle re-identification contains 10,000 images, which are captured by 326 surveillance cameras within 14 days. The resolutions of

images are distributed from 400 × 424 to 990 × 1134. VRID collects 1000 vehicle IDs (vehicle identities) of top 10 common vehicle models (**Table 1**) to reconstruct the interference with the same vehicle model in the real world. The vehicle IDs belong to the same model have very similar appearance and their differences appears in the area of the logo and accessories. Besides, each vehicle IDs contains 10 images which are in various illuminations, poses and weather condition. Dataset examples are shown in **Figure 3**.

The attributes of VRID is illustrated in **Table 2**. The vehicle model column represents the vehicle model information. The license plate column is used for the correlation of the same vehicle. The window location column shows the location of vehicle window area. The vehicle color column contains the vehicle color information. Besides, with the rich attributes of vehicles, the dataset could also be used for vehicle fine-grained recognition as well as vehicle color recognition.

3.2.2 VeRi-776

To collect high-quality videos in real-world surveillance scene, we select 20 cameras deployed along a circular road of a 1.0 km^2 area as shown in **Figure 4**.

Vehicle model	Vehicle IDs	Total images
Audi_A4	100	1000
Honda_Accord	100	1000
Buick_Lacrosse	100	1000
Volkswagen_Magotan	100	1000
Toyota_Corolla_I	100	1000
Toyota_Corolla_II	100	1000
Toyota_Camry	100	1000
Ford_Focus	100	1000
Nissan_Tiida	100	1000
Nissan_Sylphy	100	1000

Table 1.
The 10 vehicle models in the dataset.

Figure 3.
Vehicle samples selected from the VRID-1 dataset.

Image_ID	Vehicle model	License plant number	Window location	Color
IDs_1	Toyota_Corolla	License_1	X1, Y1, X2, Y2	Yellow
IDs_12	Toyota_Corolla	License_2	X1, Y1, X2, Y2	Black
IDs_1000	Honda_Accord	License_10	X1, Y1, X2, Y2	White

Table 2.
The attributes of VRID.

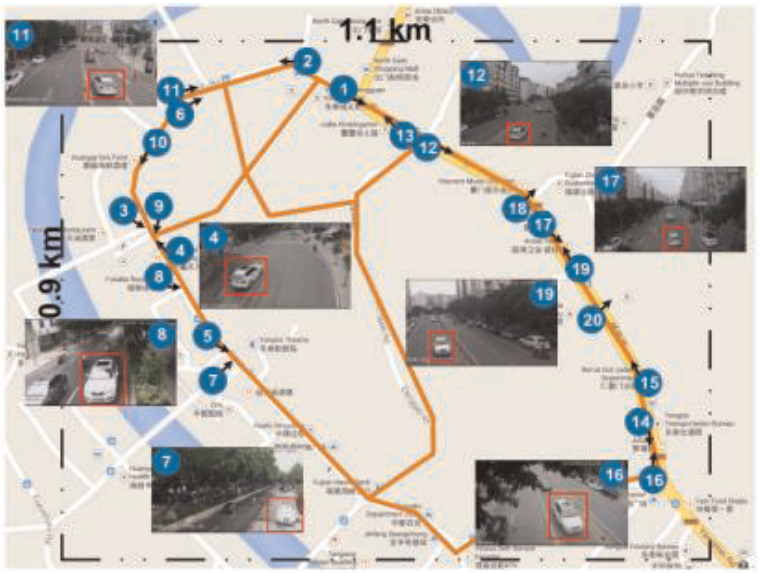

Figure 4.
The urban surveillance environments and cameras distribution for the VeRi dataset.

Figure 5.
Vehicle samples selected from the VeRi dataset.

The scenes of the cameras include two-lane roads, four-lane roads, and crossroads. All cameras are set to 1920 × 1080 resolution and 25 fps. The cameras are deployed with arbitrary orientations and tilt angles. Besides, there are overlaps for part of the cameras.

The VeRi dataset is collected with 20 cameras in real-world traffic surveillance environment. A total of 776 vehicles are annotated. Two hundred vehicles are used for testing. The remaining 576 vehicles are for testing. There are 11,579 images in the test set, 1678 images as queries and 37,778 images in the training set. Each vehicle is captured by at least two cameras. One advantage of this data set is that the camera ID and timestamp (frame ID) are reserved with tracks for further annotation. Dataset examples are shown in **Figure 5**.

4. General technical route

In deep learning method, the general technical route of object Re-ID includes three stages: data input stage, feature extraction model and distance measurement (**Figure 6**).

4.1 Data input

Data input mainly refers to feeding data to feature extraction model, and the commonly used data type in object Re-ID is three-channel image. In this part, we do not describe the input data, but mainly introduce data augmentation. In the training stage of deep learning model, insufficient data often leads to the situation that the model cannot converge or overfit. In order to avoid this situation, data augmentation is one of the solutions. Common operations for data augmentation are as follows:

- Color Jittering. Color data enhancement, such as Change image brightness, saturation, contrast and so on.
- Random Scale. Randomly change the original size of the image.
- Horizontal/vertical flip. Flip the original image horizontally or vertically.

In the data input stage, we need to pay attention to not only the data amplification, but also, in some special cases such as contrast loss or triplet loss model, we may need to construct the image pair or triplet sample in advance. Due to the limitation of GPU memory, it is impossible to input a batch of data includes all images, so it is possible that there is no negative sample which might result in the failure of image pair or triplet sample construction, at the same time, due to the large number of target individuals in the re-identification problem, the imbalance between positive and negative samples is very likely to exist, which easily leads to the unscientific network model trained. Therefore, we need to set some rules in the data input stage to correctly construct these image pairs or triplet samples.

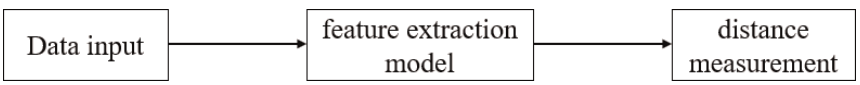

Figure 6.
General technical route of object re-identification.

4.2 Feature extraction model

The core of object Re-ID algorithm is feature extraction model, the effectiveness of the whole algorithm is also almost determined by this part. In other words, the essence of Re-ID is to compare the similarity or distance between the features extracted from two images. Image features mainly include color feature, texture feature, shape feature and spatial relationship feature. Feature extraction is a concept in computer vision and image processing. It refers to the use of computer to extract image information to determine whether each image pixel belongs to an image feature. Features are the best way to describe patterns, and we often think that each dimension of a feature can describe a pattern from a different perspective. Ideally, the dimensions are complementary and complete. In the field of image recognition or image Re-ID, traditional methods of feature extraction include Histogram of Oriented Gradient (HOG), scale-invariant feature transform (SIFT), Speeded Up Robust Features (SURF), Local Binary Pattern (LBP) and so on; the deep learning methods of feature extraction include Convolution Neural Network (CNN), Recurrent Neural Network (RNN) and so on. We present a feature extraction method in detail in both traditional and deep learning methods.

4.2.1 Histogram of oriented gradient (HOG)

The essence of HOG feature extraction is to constitute features by computing and statistics the histogram of gradient direction in the local area of the image. Hog feature combined with SVM classifier has been widely used in image recognition, especially in pedestrian detection, which has achieved great success. How to extract HOG feature? Firstly, the image is divided into small connected regions, which are called cell units. Then the direction histogram of the gradient or edge of each pixel in the cell is collected. Finally, these histograms can be combined to form a feature descriptor.

4.2.2 Convolution neural network (CNN)

It is a kind of feedforward neural network with deep structure including convolution calculation. A convolutional neural network contains three types of neural network layers: convolutional layer, pooling layer and fully connection layer. As is shown in **Figure 7**.

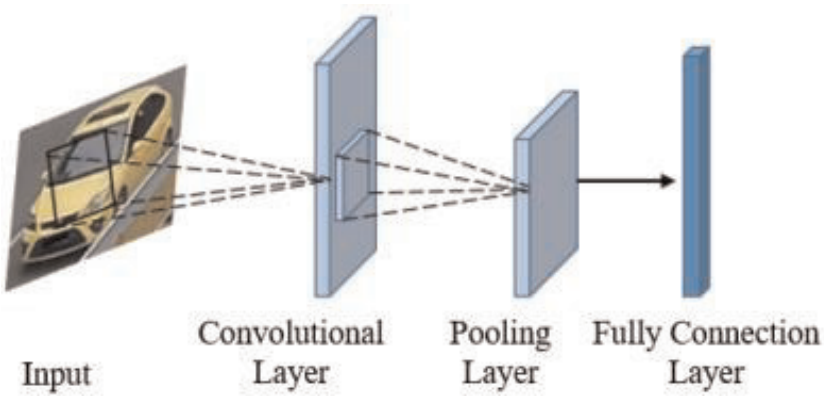

Figure 7.
Basic structure diagram of convolutional neural network.

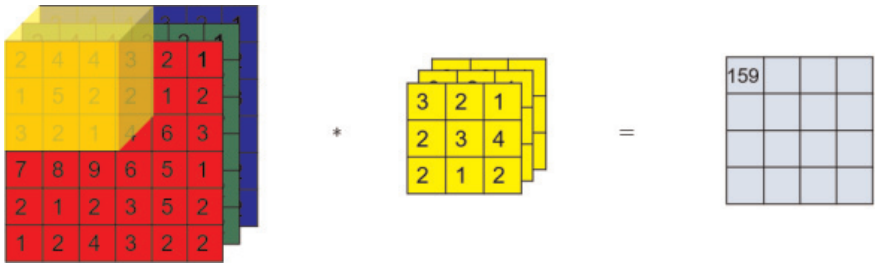

Figure 8.
Convolution diagram.

4.2.2.1 Convolutional layer

The convolution layer is mainly used for learning the feature representation of input data. The convolution layer is composed of multiple convolution kernels, and the convolution operation is carried out on the input image to calculate different feature maps.

In general, the input data is RGB image, as shown in **Figure 8**. If the color image is 6*6*3, the three refers to three color channels, and the convolution operation is carried out with a 3*3*3 convolution kernel, corresponding to the red, green and blue channels. Take the 27 numbers in turn, multiply them by the Numbers in the corresponding red, green and blue channel, and then add them all up to get the first number in the output of the feature graph.

The convolution layer principle is shown in equation:

$$x_j^l = f\left(\sum_i x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

where $f(*)$ is activation functions; x_j^l denotes the $j-th$ feature map of output layer l, x_i^l is the i $-th$ feature map of the layer l; K_{ij}^l represents the convolution kernel of the i $-th$ feature graph of the current input layer and the $j-th$ feature graph of the output layer on the layer l; b_j^l is the bias term of the $j-th$ feature graph in the layer l.

4.2.2.2 Pooling layer

Pooling layer is often used in the convolutional network to reduce the size of the model, improve the computational speed, and improve the robustness of extracted features. Pooling operation can maintain the invariance of translation, rotation and scale. Common pooling layer operations are averaging and pooling. The maximum pooling operation is as shown in **Figure 9**. The input of 4*4 is divided into different regions. For the output of 2*2, each element output is the maximum element value in its corresponding color region.

4.2.2.3 Fully connection layer

Each node of the fully connection layer is connected to all nodes of the previous layer to integrate the features extracted from the previous layer. Due to its fully connected nature, the general fully connected layer also has the most parameters. The full join layer act as a mapping of the learned "distributed feature representation" into the sample tag space. It's essentially a linear transformation from one

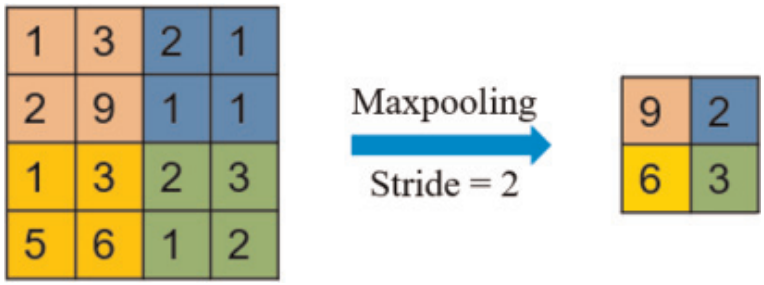

Figure 9.
Max pooling diagram.

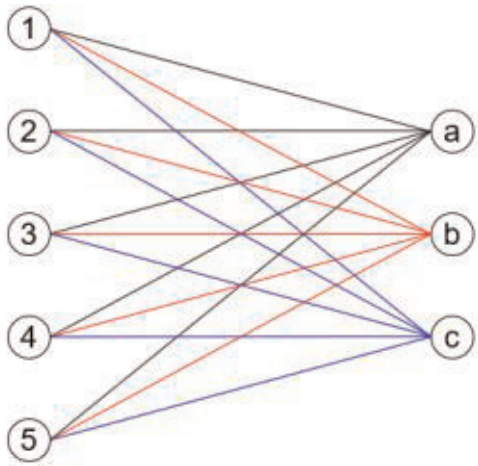

Figure 10.
Fully connection layer.

eigenspace to another eigenspace. Any dimension of the target space is affected by every dimension of the source space. In CNN, the full connection is often found in the last few layers, which is used to make a weighted sum of the features designed before. The schematic diagram of the entire connection layer is shown in **Figure 10**.

4.3 Distance measurement

After feature extraction, we need to compare the distance between the query image and all images in the retrieval set, and there are many ways you can measure the difference between two features, It is divided into distance measure (such as, Euclidean distance, Manhattan Distance etc.) and similarity measure (such as, Cosine Similarity, Jaccard Coefficient, etc.).

Distance measure is used to measure the distance of an individual in space, the greater the distance, the greater the difference between individuals. Similarity measurement is to calculate the degree of similarity between individuals. Contrary to distance measurement, the smaller the value of similarity measurement is, the smaller the similarity between individuals is, and the greater the difference is. Therefore, we can judge which images are more likely to be the same individual by the value of the difference between image features.

5. One case of person Re-ID

Person Re-ID is a technology that uses computer vision technology to judge whether there is a specific person in the image or video sequence. It is widely regarded as a sub-problem of image retrieval. Given a monitor person image, retrieve the image of the row of people across the device. It aims to make up for the visual limitations of the current fixed cameras, and can be combined with person detection and pedestrian tracking technology, which can be widely used in intelligent video monitoring, intelligent security and other fields. In this section, we show a classic person Re-ID algorithm Part-based Convolutional Baseline (PCB) [43].

5.1 Structure of PCB

PCB can take any network without hidden fully-connected layers designed for image classification as the backbone, e.g., Google Inception and ResNet. Original paper employs ResNet50 as the backbone network to reproduce the PCB algorithm.

The structure of PCB illustrated in **Figure 11**. The input image goes forward through the stacked convolutional layers from the backbone network to form a 3D tensor T. PCB replaces the original global pooling layer with a conventional pooling layer, to spatially down-sample T into p pieces of column vectors g. A following 1×1 kernel-sized convolutional layer reduces the dimension of g. Finally, each dimension-reduced column vector h is input into a classifier, respectively. Each classifier is implemented with a fully-connected (FC) layer and a sequential Softmax layer. During training, each classifier predicts the identity of the input image and is supervised by Cross-Entropy loss. During testing, either p pieces of g or h are concatenated to form the final descriptor of the input image.

5.2 Experimental results

5.2.1 Dataset

The original paper tested this algorithm on person Re-ID dataset Market-1501. The Market-1501 dataset contains 1501 identities observed under six camera viewpoints, 19,732 gallery images and 12,936 training images detected by DPM.

5.2.2 Performance comparison

It compares PCB and PCB + RPP with state of the art. Comparisons on Market-1501 are detailed in **Table 3**. PCB + RPP get mAP = 81.6% and Rank-1 = 93.8% for Market-1501, setting new state of the art on this dataset. All the results are achieved under the single-query mode without re-ranking. Reranking methods will further

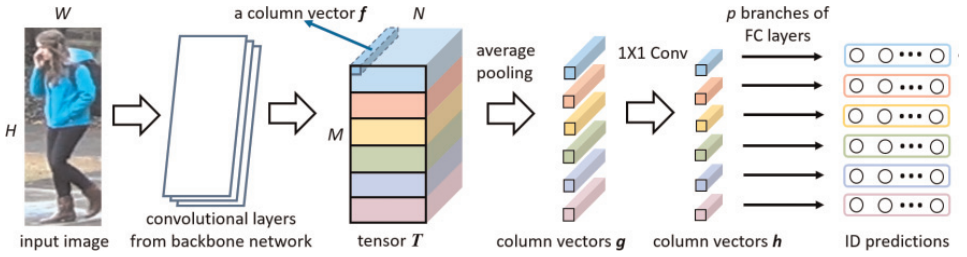

Figure 11.
Structure of PCB [43].

Methods	Rank-1	Rank-5	Rank-10	mAP
KLFDA	46.5	71.1	79.9	—
Triplet Loss	84.9	94.2	—	69.1
DML	87.7	—	—	68.8
MultiScale	88.9	—	—	73.1
GLAD	89.9	—	—	73.9
PCB	92.3	97.2	98.2	77.4
PCB + RPP	**93.8**	**97.5**	**98.5**	**81.6**

Table 3.
Comparison of the proposed method with the art on Market-1501.

boost the performance especially mAP. For example, when "PCB + RPP" is combined with the reranking method, mAP and Rank-1 accuracy on Market-1501 increases to 91.9 and 95.1%, respectively.

6. One case of vehicle Re-ID

Considering the spatiotemporal logic of vehicle driving process, we present a vehicle re-identification (Re-ID) algorithm based on multi-camera data's spatiotemporal information and joint learning mechanism without license plate. The algorithm is divided into feature extraction and spatiotemporal re-rank. In the feature extraction stage: on the basis of convolutional neural network (CNN), triplet loss and Softmax loss were used for joint training to model a feature extractor and calculate the feature distance measurement matrix between query image and retrieval set images. In the spatiotemporal re-rank stage: we calculate the spatiotemporal distance matrix and fuse the spatiotemporal distance with the normalized feature distance metric. The final distance measurement matrix is sorted to obtain the vehicle re-identification result. Extensive experiments were carried out on the benchmark datasets "VeRi" to verify the effectiveness of the proposed method and the result have shown that the proposed algorithm outperforms the state-of-the-art approaches for vehicle Re-ID.

6.1 Mathematical principles of joint learning

The architecture of proposed algorithm is illustrated in **Figure 12**. The algorithm is divided into two steps: feature extraction and spatiotemporal re-rank. In the feature extraction phase, triplet loss and Softmax loss are integrated for joint training, triplet loss is used to calculate the distance of the sample features, increasing the distance between the anchor and negative sample, reducing the distance between the anchor and the positive sample. Softmax loss performs label-level

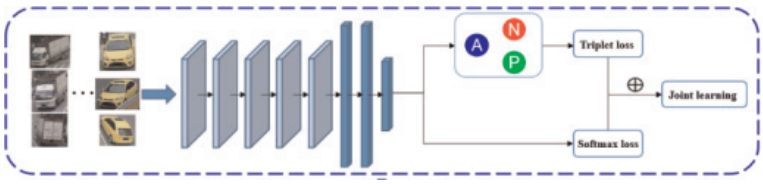

Figure 12.
The proposed algorithm for vehicle Re-ID.

supervision and constraint on the feature extraction network. In the spatiotemporal re-rank stage, calculating the spatiotemporal distance between images, and re-rank the retrieval results by merging the spatiotemporal distance and the feature distance.

6.1.1 Triplet loss

In order to learn high discriminative features from images to Euclidean space, where the distance can measure the discrepancy between two images. The idea of learning to rank has gradually been applied to many fields, such as face recognition [12], person Re-ID, and so on. One of the important steps in learning to rank is to find a good similarity function, and triplet loss is a very broad one. In the calculation of the triple loss, the feed data includes an anchor, a positive sample, and a negative sample, and the sample similarity calculation is realized by optimizing the distance between the anchor and the positive sample being smaller than the distance between the anchor and the negative sample. We suppose $T = \{x_i | i = 1, 2...m\}$ denotes the training set, where x_i is the $i - th$ image in the training set and m denotes the total amount of training images. For an image triplet $\{x_i^a, x_i^p, x_i^n\}$, where x_i^a denotes an anchor, x_i^p denotes a positive of the same class as the anchor, x_i^n denotes a negative of a different class as the anchor, the triplet loss is calculated as Eq. (2).

$$L_{triplet} = \sum_{i}^{m} \max\left(0, \left\|f\left(x_i^a\right) - f\left(x_i^p\right)\right\|_2^2 - \left\|f\left(x_i^a\right) - f\left(x_i^n\right)\right\|_2^2 + \alpha\right) \tag{2}$$

where $f(x_i)$ denotes the embedded of the image, α denotes the parameter of expected gap between the distance of $\{x_i^a, x_i^p\}$ and $\{x_i^a, x_i^n\}$

6.1.2 Triplet sampling

This algorithm directly performs on-line triplet mining on image features, which is to compute useful triplets on the fly. For each batch of inputs, given a batch of N examples, we compute the N embeddings and we then can find a maximum of N^3 triplets. For three indices a, p, n ∈ [1, N], if examples a and p have the same label but are distinct, and example n has a different label, we say that (a, p, n) is a valid triplet. We suppose that have a batch of vehicle images as input of size N = PK, composed of P different vehicle ID with K images each. Choose the batch hard strategy: for each anchor, select the hardest positive and the hardest negative among the batch, finally we can obtain PK triplets.

$$d(a, b) = \|a - b\|^2 = \|a\|^2 - 2\langle a, p\rangle + \|b\|^2 \tag{3}$$

6.1.3 Softmax loss

We impose a strong constraint on distinguishing different vehicle label by adding Softmax loss to the loss function. The embedded obtained by CNN tend to clusters, and the embedded of same vehicle ID will be similar, so the convergence time of triplet loss will be cut down. In Softmax loss stream, each vehicle ID in the training set is considered as a category, the Softmax loss function is formulated as:

$$L_{soft\max} = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right] \tag{4}$$

where m is the total amount of classes, k is the number of training image, $1(*)$ is the indicator function (if * is true, then the value set 1, or 0), and $\theta's$ are the parameters of the final full-connection layer of the CNN.

6.1.4 Joint multiple loss

The joint learning mechanism is mainly applied to the training phase of vehicle images. After the shared images pass through the shared CNN layer, they are divided into two branch streams, one is subjected to online triplet mining for the calculation of triplet loss, and the other stream enters the Softmax layer for Softmax loss calculation. The final joint learning loss function can be formulated as:

$$L_{JL} = L_{softmax} + L_{triplet} \tag{5}$$

6.2 Experimental results

6.2.1 Dataset

In order to verify the validity of the algorithm, we conduct experiments in the latest version of the vehicle re-identification dataset "VeRi." The dataset has a total of 49,357 images, which are taken for actual road monitoring and contains various angles and various vehicle models, as shown in **Figure 13**.

It is divided into two subsets for training and testing. The train set has 576 vehicle IDs with 37,778 images and the test set has 200 vehicle IDs with 11,579 images. For the vehicle re-identification task, we divided the test set to query set (1678 images) and retrieval set (9901 images).

Figure 13.
Vehicle samples selected from the VeRi dataset.

6.2.2 Experimental setting

All of the experiments are based on the deep learning framework Tensorflow. The base network is VGG_CNN_M, and the model was pre-trained on the ImageNet. In the calculation of triplet loss, we set $\alpha = 1$, the learning rate is set to 0.001, and the mini-batch is set to 32.

In order to evaluate the effect of this algorithm objectively, we set up two algorithms to compare with the method proposed to verify that the improvement of the algorithm. These algorithms are: (1) VGG + Softmax loss; (2) VGG + Triplet loss; (3) VGG + Softmax loss + Triplet loss (our method). All of network based on VGG16, "Softmax loss" denotes use Softmax loss to train the network, and "Triplet loss" denotes use triplet loss to train the network. At the same time, we also make comparison our experiment results with some state-of-the-art algorithms on the same dataset "VeRi."

6.2.3 Performance comparison on VeRi dataset

We conduct the experiment as described in experimental setting, and use cumulative matching curve (CMC), HIT@1, HIT@5 as metrics to evaluate the performance. In our method, S, T denote using Softmax loss and using triplet loss respectively. **Table 4** and **Figure 14** illustrate the performances of the proposed methods and some state-of-the-art algorithms in vehicle Re-ID field.

The results show that the proposed method "VGG + S + T" achieves the best results, the HIT@1 and HIT@5 hit 89.75 and 95.05% respectively. It is obvious that the CNN-based method has a significant improvement over the handcraft feature-based approach when compare BOW-CN and LOMO algorithm with other algorithms based on CNN feature. Compared with "VGG + S" which only utilizes Softmax loss, our method has much better results, improving 16.81% in HIT@1 and 11.38% in HIT@5. Compared with "VGG + T" which only utilizes triplet loss, our method makes improvement about 16.81% in HIT@1 and 8.67% in HIT@5. Compare to "FACT + Plate-SNN + STR" which additionally utilizes license plate information (Plate-SNN) and spatiotemporal relation (STR), our method improves 28.31% in HIT@1 and 16.27% in HIT@5. In summary, the proposed algorithm is feasible in vehicle re-identification task, and achieves outstanding results compared to other algorithms.

Method		HIT@1	HIT@5
BOW-CN		33.91	53.69
LOMO		25.33	46.48
ABLN		58.14	74.41
FACT + Plate-SNN+ STR		61.44	78.78
VAMI		77.03	90.92
JFSDL		82.90	91.60
This method	VGG+ S	72.94	83.67
	VGG + T	72.94	86.83
	VGG + S + T	**89.75**	**95.05**

Table 4.
Comparison of the proposed method with the art on VeRi.

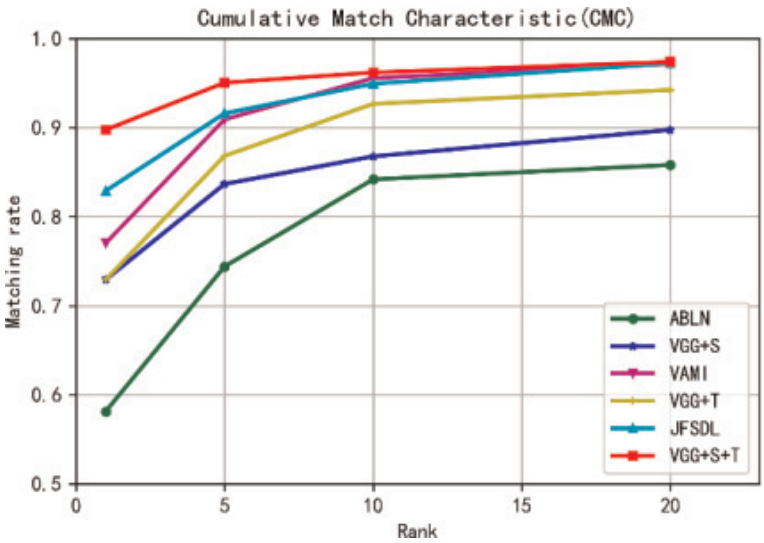

Figure 14.
The CMC curves on VeRi.

7. Summary

This chapter mainly introduces the concept of object Re-ID and two core applications: person Re-ID and vehicle Re-ID. In this chapter, the definitions of person Re-ID and vehicle Re-ID are given, some research methods of the two applications are reviewed, and the commonly used public data sets are described in detail.

In this chapter, the general process of object Re-ID by deep learning method is given, and the data input, feature extraction network structure, distance measurement and other parts are described in detail. At the same time, two examples are given to illustrate the algorithm in detail and experiment comparison. Person Re-ID refers to the network structure and experimental results of PCB algorithm [43]. Vehicle Re-ID is introduced in detail in terms of feature extraction and measurement calculation. The influence of parameters in the deep learning method is illustrated through the analysis of experimental results, and the evaluation comparison is given.

These can help relevant researchers to understand the context of technology, the general implementation process, as well as important parameters and evaluation indicators in this field, so that they can quickly start relevant research.

The object Re-ID is the basis of realizing cross-camera tracking. Person and vehicles are just two typical applications. In the future, with the gradual solution of the following problems, we will have a more extensive application:

- High-quality standard database is important to generalization performance of Re-ID algorithm. The database should be more suitable for the real environment and including different and varying scenes.

- Deep networks have poor interpretability. Although the deep learning method has achieved good performance in Re-ID tasks, few studies have shown which information has a greater impact on Re-ID behind the continuous improvement in accuracy.

- At present, most methods are carried out under the prior condition that object has been detected, but this requires a very robust detection model. We need to combine object Re-ID with object detection, which is more in line with practical application requirements.

- The research should focus on semi-supervised, unsupervised and transfer learning methods. The collected data are limited after all, and the cost of labeling data is also very high. Therefore, although the semi-supervised and unsupervised learning methods may not be as good as the supervised learning methods in terms of performance, they are valuable.

Acknowledgements

This work is supported by the National key Research and Development Program of China under Grant No. 2018YFB1601101 of 2018YFB1601100 and National Natural Science Foundation of China under Grant No. U1611461.

Author details

Xiying Li[1,2,3]* and Zhihao Zhou[1,2,3]

1 School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou, Guangdong, People's Republic of China

2 Guangdong Provincial Key Laboratory of Intelligent Transportation System, Guangzhou, Guangdong, People's Republic of China

3 Laboratory of Video and Image Intelligent Analysis and Application Technology, Ministry of Public Security, Guangzhou, Guangdong, People's Republic of China

*Address all correspondence to: stslxy@mail.sysu.edu.cn

References

[1] Luo H, Wei J, Xing F, Si-Peng Z. A survey on deep learning based person re-identification. Acta Automatica Sinica. 2019. DOI: 10.16383/j.aas.c180154

[2] Sunderrajan S, Manjunath BS. Context-aware hypergraph Modeling for Re-identification and summarization. IEEE Transactions on Multimedia. 2016;**18**(1):51-63

[3] Geng M, Wang Y, Xiang T, Tian Y. Deep transfer learning for person reidentification. arXiv preprint arXiv: 1611.05244; 2016

[4] Lin Y, Zheng L, Zheng Z, Wu Y, Yang Y. Improving person re-identification by attribute and identity learning. arXiv preprint arXiv: 1703.07220; 2017

[5] Varior RR, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification. In: European Conference on Computer Vision; Springer; 2016. pp. 791-808

[6] Shengyong D, Liang L, Guangrun W, et al. Deep feature learning with relative distance comparison for person reidentification. Pattern Recognition. 2015;**48**(10):2993-3003

[7] Chen W, Chen X, Zhang J, Huang K. Beyond triplet loss: A deep quadruplet network for person re-identification. 2017;**1**:1320-1329

[8] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person reidentification. arXiv preprint arXiv: 1703.07737; 2017

[9] Qiqi X, Hao L, Chi Z. Margin sample mining loss: A deep learning based method for person re-identification. arXiv preprint arXiv: 1710.00478; 2017

[10] Longhui W, Shiliang Z, Hantao Y, et al. GLAD: Global-local-alignment descriptor for scalable person re-identification. IEEE Transactions on Multimedia. 2019;**21**(4):986-999

[11] Chi S, Jianing L, Shiliang Z, et al. Pose-driven deep convolutional model for person reidentification. IEEE International Conference on Computer Vision (ICCV). 2017;**1**:3980-3989

[12] Zheng L, Huang Y, Lu H, Yang Y. Pose invariant embedding for deep person reidentification. arXiv preprint arXiv:1701.07732; 2017

[13] Xuan Z, Hao H, Xing F, et al. AlignedReID: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184

[14] Liu H, Jie Z, Jayashree K, et al. Video based person re-identification with accumulative motion context. IEEE Transactions on Circuits and Systems for Video Technology. 2018;**28**(10): 2788-2802

[15] Mazzeo PL, Spagnolo P, D'Orazio T. Object tracking by non-overlapping distributed camera network. In: International Conference on Advanced Concepts for Intelligent Vision Systems; 2009;**5807**:516-527

[16] Mazzeo P, Giove L, Moramarco GM, Spagnolo P, Leo M. HSV and RGB color histograms comparing for objects tracking among non-overlapping. In: 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS); IEEE; 2011. pp. 498-503

[17] Tian Y, Dong H-h, Jia L-m, et al. A vehicle re-identification algorithm based on multi-sensor correlation. Journal of Zhejiang University-Science C (Computers & Electronics). 2014; **15**(5):372-382

[18] Coifman B. Vehicle reidentification and travel time measurement, part II: Uncongested freeways and the onset of congestion. Journal of Transportation Engineering. 2003;**129**(5)

[19] Rios-Cabrera R, Tuytelaars T, Van Gool L. Efficient multi-camera vehicle detection, tracking, and identification in a tunnel surveillance application. Computer Vision and Image Understanding. 2012;**116**(6):742-753

[20] Liu X, Liu W, Ma H, et al. Large-scale vehicle re-identification in urban surveillance videos. In: 2016 IEEE International Conference on Multimedia and Expo (ICME); Barcelona, Spain: IEEE; 2016

[21] Liu H, Tian Y, Wang Y, et al. Deep relative distance learning: tell the difference between similar vehicles. In: Computer Vision and Pattern Recognition; Las Vegas, USA: IEEE; 2016. pp. 2167-2175

[22] Li Y, Li Y, Yan H, et al. Deep joint discriminative learning for vehicle re-identification and retrieval. In: 2017 IEEE International Conference on Image Processing (ICIP); Beijing, China: IEEE; 2017

[23] Zhang Y, Liu D, Zha ZJ. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME); 2017. pp. 1386-1391

[24] Marin-Reyes PA, Bergamini L, Lorenzo-Navarro J, et al. Unsupervised vehicle re-identification using triplet networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE; 2018

[25] Liu X, Liu W, Mei T, et al. Provid: Progressive and multimodal vehicle reidentification for large-Scale Urban surveillance. In: IEEE Transactions on Multimedia; 2017. p. 99

[26] Na J, Yue X, Zhou Z, et al. Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking. In: ICIP; 2018

[27] Yantao S, Tong X, Hongsheng L, et al. Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals. In: 2017 IEEE International Conference on Computer Vision (ICCV); IEEE; 2017

[28] Li X, Zhou Z, Qiu M. A vehicle re-identification algorithm based on component fusion feature. Computer Engineering. DOI: 10.19678/j.issn.1000-3428.0052284

[29] Zhou Z, Li X, Qiu M. A car face parts detection algorithm based on faster R-CNN. In: 18th COTA International Conference of Transportation Professionals: Intelligence, Connectivity, and Mobility (CICTP 2018); 2018

[30] Liang D, Yan K, Wang Y, et al. Deep hashing with multi-task learning for large-scale instance-level vehicle search. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW); IEEE Computer Society; 2017

[31] Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). Rio de Janeiro: IEEE; 2007. pp. 1-7

[32] Hirzer M, Beleznai C, Roth PM, et al. Person re-identification by descriptive and discriminative classification. In: Proceedings of Scandinavian Conference on Image Analysis. Berlin, Heidelberg: Springer; 2011. pp. 91-102

[33] Li W, Zhao R, Xiao T, et al. DeepReID: Deep filter pairing neural network for person re-identification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; Columbus, OH, USA: IEEE; 2014. pp. 152-159

[34] Zheng L, Shen L, Tian L, et al. Scalable person re-identiflcation: A benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision; Santiago, Chile: IEEE; 2015. pp. 1116-1124

[35] Tong X, Shuang L, Bochao W, et al. End-to-end deep learning for person search. arXiv preprint arXiv: 1604.01850; 2016

[36] Zheng L, Bie Z, Sun Y, et al. Mars: A video benchmark for large-scale person reidentification. In: Proceedings of European Conference on Computer Vision; Cham: Springer; 2016. pp. 868-884

[37] Ristani E, Solera S, Zou R, Cucchiara R, et al. Performance measures and a data set for multi-target, multicamera tracking. In: Proceedings of European Conference on Computer Vision; Cham: Springer; 2016. pp. 17-35

[38] Wu A, Zheng WS, Yu HX, et al. RGB-infrared cross-modality person re-identification. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; Venice, Italy: IEEE; 2017. pp. 5390-5399

[39] Song G, Leng B, Liu Y, et al. Region-based quality estimation network for large-scale person reidentification. In: Proceedings of Association for the Advancement of Artificial Intelligence; New Orleans: AAAI; 2018

[40] Wei L, Zhang S, Gao W, et al. Person transfer GAN to bridge domain gap for person re-identification. arXiv: 1711.08565; 2018

[41] Li J, Zhang S, Wang J, et al. LVreID: Person re-identification with long sequence videos. arXiv preprint arXiv: 1712.07286; 2017

[42] Li X, Yuan M, Jiang Q, et al. VRID-1: A basic vehicle re-identification dataset for similar vehicles. In: IEEE, International Conference on Intelligent Transportation Systems; IEEE; 2017. pp. 1-8

[43] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). 2018