Research Article

A Vehicle Reidentification Algorithm Based on Double-Channel Symmetrical CNN

Lijun Yang[1] **and Tangsen Huang** [ORCID][2]

[1]*Shaanxi Xueqian Normal University, Xi'an 710100, China*
[2]*School of Electronics and Information Engineering, Hunan University of Science and Engineering, Yongzhou 425199, China*

Correspondence should be addressed to Tangsen Huang; huangtangsen@163.com

It has become a challenging research topic to accurately identify the vehicles in the past from the mass monitoring data. The challenge is that the vehicle in the image has a large attitude, angle of view, light, and other changes, and these complex changes will seriously affect the vehicle recognition performance. In recent years, the convolutional neural network (CNN) has achieved great success in the field of vehicle reidentification. However, due to the small amount of vehicle annotation in the dataset of vehicle reidentification, the existing CNN model is not fully utilized in the training process, which affects the ability to identify the deep learning model. In order to solve the above problems, a double-channel symmetric CNN vehicle recognition algorithm is proposed by improving the network structure. In this method, two samples are taken as input at the same time, in which each sample has complementary characteristics. In this case, with limited training samples, the combination of inputs will be more diversified, and the training process of the CNN model will be more abundant. Experiments show that the recognition accuracy of the proposed algorithm is better than other existing methods, which further verifies the effectiveness of the proposed algorithm in this study.

1. Introduction

In recent years, the society pays more and more attention to the public security problem, and the monitoring equipment is more and more popular. A large number of surveillance cameras are used in crowded places prone to public security incidents, such as traffic intersections, parks, large shopping malls, stations, and airports. The emergence of surveillance cameras has brought great convenience to the case detection of public security organs, such as suspected vehicle chase, cross-scene vehicle search, abnormal event detection, and so on [1, 2]. A large number of surveillance cameras form a huge network of surveillance. Although the monitoring system has developed rapidly, it has brought great challenges to the management and analysis of monitoring data [3, 4]. At present, the monitoring system mostly adopts the method of real-time camera and human participation to monitor. The massive monitoring data is a big problem for the personnel who are in charge of monitoring the video. There are two

reasons: (1) the monitoring system generates data in real time, resulting in a huge amount of data; (2) the real-time monitoring data records a scene with random changes, and it is difficult for the monitor staff to pay attention for a long time during the long-time observation. It can be seen that this kind of monitoring mechanism with human participation is no longer applicable to the management and analysis of monitoring data. However, the emergence of vehicle reidentification technology overcomes the deficiency in the supervision mechanism of human participation.

In recent years, deep learning models represented by the convolutional neural network (CNN) have achieved great success in the field of computer vision. At the same time, CNN also led the research in the field of vehicle reidentification. Compared with the traditional vehicle reidentification method designed by hand, the vehicle reidentification method based on CNN can overcome the complex changes of vehicles more effectively and achieve higher performance. However, vehicle reidentification is different from other

computer vision tasks because it is very difficult to annotate the vehicles, resulting in a small amount of vehicle annotation in the existing dataset. On the limited training set of the current picture training set, the training of the existing single-channel CNN model will make the training process of the CNN model insufficient. In order to make the combination of input images more diversified, multiple combinations of images can be used as input to fully train the CNN network. At the same time, the recognition rate will be improved because the double-channel CNN network can input more features.

This study attempts to design a double-channel symmetrical CNN structure for vehicle reidentification by improving the network structure. In this double-channel structure, two samples are input at a time. At this time, compared with the previous single-channel CNN model, the input combination form of this double-way CNN model is more diversified, which is suitable for the deep learning model with stronger ability of learning and obtaining discrimination.

2. Related Works

The task of vehicle reidentification [5, 6] is to study how to accurately identify the vehicle that has appeared in a particular occasion in the mass monitoring data, in which the monitoring data are mainly image data. The challenge of the task is that the vehicle in the image has a large attitude, angle of view, and other complex changes. In addition, during the shooting process, different lighting will also make the appearance of the vehicle change greatly. The above changes will seriously affect the performance of vehicle recognition. At present, the research on target reidentification mainly focuses on the field of pedestrian reidentification [7–10] and is rarely applied to other targets. Since 2015, a small number of scholars have tried to enter the field of vehicle reidentification, but they can only be applied to images of the same scale and angle, with weak robustness to environmental changes or based on small datasets.

In order to improve re-ID capability, some methods utilize additional attribute information such as the model/type and color to guide vision-based representation learning [11]. For example, [12] introduced a two-branch retrieval pipeline to extract differences between models and instances. Yan et al. [13] studied the multiparticle relationship of vehicles with multilevel attributes. Other works study temporal and spatial associations, which derive additional benefits from the topological information of the camera [14]. In addition, some methods use GAN [15] to generate images from the required viewpoint, so as to achieve viewpoint alignment. It can be said that these works solve the problem of viewpoint change through viewpoint alignment.

In addition, [16] claimed that in addition to the dataset used for training, the features made by traditional handwork were easy to generate deep features, so the two features were combined to achieve an improved representation. Liu et al. [17] used a multimodal finite element analysis including visual features, license plate, camera position, and other contextual information in a coarse to fine vehicle retrieval framework. In order to enhance training data and achieve robust training, [18] used the generated countermeasure network to synthesize vehicle images with different directions and appearance changes. Zhou and Shao [15] through the attention model of advertising learning and visual perception, the visual perception representation of vehicle re-ID is learned. Zhang et al. [19] proposed an improved joint optimization of three-one-loss execution and an auxiliary classification loss as a regularization to represent the in-sample variance.

3. Single-Channel CNN Structure

The single-channel CNN structure is introduced in this section first; then, double-channel CNN structure is detailed in the next section.

In the training set of vehicle reidentification, the single-channel CNN model based on identification is used for learning, so that the deep learning model obtained after training can distinguish different vehicles. Based on the existing classic CNN model, all the convolutional layers and full connection layers in the AlexNet [20] and ResNet-50 [21] models are used. The default parameters provided in literature [20, 21] are adopted, and the output of the last full connection layer is modified to be the total number of different vehicles in the vehicle reidentification training set. The CNN model of the single-channel method is the fine-tuning of the pretraining model obtained on the ImageNet dataset [22], at which time the convergence rate of the CNN model is faster. Especially, in the case that the scale of the vehicle reidentification training set is not very large, this training strategy is more effective and achieves the purpose of distinguishing different vehicles.

The network training process is described as follows. The vehicle reidentification training set is recorded as $D = \{x_i, y_i\}_{i=1}^{N}$, the vehicle image is x_i, and the identity (ID) is y_i. The vehicle image is first processed to a size of 256×256 pixels and then randomly cropped to a fixed size (AlexNet is 227×227 pixels and ResNet-50 is 224×224 pixels). The processed vehicle image is sent to the data layer of the CNN model as an input to the network. The goal of network training is to get a deep learning model M through deep learning. It is equivalent to mapping: $f(x, \theta) \in R^C$, where θ represents the parameters of each layer in the CNN model. In the process of each minibatch iteration, the parameter Q is updated using the random gradient descent (SGD) algorithm. In the t iteration, the current parameter Q is updated as follows

$$\theta_{t+1} = \theta_t + \gamma \cdot \frac{1}{|Dt|} \sum_{(x,y \in D_t)} \nabla_{\theta_t}[l(x, y)], \tag{1}$$

where γ is the learning rate, D_t is a set of minibatch samples taken randomly from D, ∇ is the gradient operation, and l is the loss function, which is the softmax loss function. The softmax loss function acts as a supervisory signal to guide the network training process. As the training process progresses, the value of the loss function gradually decreases. At this point, the trained network is convergent.

In the process of vehicle recognition, the deep learning model M obtained from network training is used as the feature extractor. The middle layer of the probe set and gallery set of the vehicle image is processed, and the response of the middle layer is extracted as the feature. AlexNet is set as the response of the FC7 layer, and ResNet-50 is set as the response of the Pool5 layer. On the basis of image features, cross-camera retrieval is performed, that is, the distance of image features between the samples in the probe set and the gallery set is calculated. The distance is sorted, and the final vehicle rerecognition performance is evaluated against the sorted list.

4. Proposed Double-Channel Symmetric CNN Structure

The vehicle recognition method proposed by the double-channel symmetric CNN structures is described in this section. The overall structure of the model is shown in Figure 1 (taking the AlexNet model as an example). Compared with the existing single-channel CNN model, the proposed double-channel symmetric CNN model inputs two samples at same time, and the input combination forms are more diversified. Each middle layer has the same structure and can be considered symmetrical but does not share parameters with each other. By connecting the last fully connected layer in the double-channel model, each layer in the double-channel model interacts with each other and promotes each other, which can be considered as complementary.

The goal of the network training process of the identification model is to learn an optimal mapping for a given training set, so that the prediction results of vehicles are closer to their real identity (ID). On the one hand, the richer the sample in the training set, the stronger the generalization ability of the model obtained. On the other hand, for a particular vehicle, the difference in appearance is more obvious because it is a vehicle image collected under cross-camera. By combining different vehicle images within a specific vehicle, the samples can complement each other and narrow the differences in appearance. Therefore, the designed structure is more suitable for the deep learning model with stronger discrimination ability to be learned, so as to improve the performance of vehicle rerecognition.

In the proposed double-channel symmetry CNN structure, two vehicle images are input at the same time each time, and the two images belong to the same vehicle. These sample pairs are a pairwise combination of all samples corresponding to the same carrier in a full permutation form. The preprocessing of vehicle image before sending to the network data layer is consistent with the single-channel method. Each convolution layer and the full connection layer have the same structure and settings, and each CNN model is fine-tuned by the pretraining model obtained on the ImageNet dataset. An example of the AlexNet model is showed in Figure 1. The full connection layers of FC6 and FC7 in each road are, respectively, connected with their convolution layers. The full connection layers of FC7 in the two channels are connected in series, denoted as FC7_concat. $N_1 = 4096$ dimensions; $N_2 = 8192$ dimensions. Three fully connected layers (double FC7 layers and one FC7_concat layer) are, respectively, connected to the fully connected layer FC8. The number of outputs of the FC8 layer N3 is the same as the total number of vehicles in the training set.

The three softmax loss functions are used as the supervisory signals to guide the network training process, and the sum of the three loss functions is used as the network loss. If the intermediate skeleton of the proposed two complementary symmetrical CNN structures is replaced by the ResNet-50 network, since the last layer of the network is the pool layer Pool5 and not the full connection layer FC7, then Pool5 is used instead of FC7, and the connected Pool5 layer can be denoted as Pool5_concat, where $N_1 = 2048$ dimensions and $N_2 = 4096$ dimensions. The network training strategy and the process of the proposed double-channel symmetric CNN structure are the same as the single-channel method.

The process of vehicle reidentification is to use the deep learning model obtained in the process of network training as a feature extractor. It extracts the response of the middle layer (AlexNet is the response of the FC7_concat layer and ResNet-50 is the response of the Pool5_concat layer) as the feature representation of the vehicle image in the probe set and the gallery set. On the basis of image features, the cross-camera search is performed to calculate the distance between the image features in the probe set and the gallery set, and the distance is sorted. The final vehicle recognition performance is evaluated according to the ranking list.

5. Experimental Results and Analysis

5.1. Dataset Construction. The test vehicle dataset is collected by 4 different intersection monitoring platforms, and the installation location is shown in Figure 2. The same angle video is taken every 2 hours at an interval angle of 30°, and a total of 7 angle mp4 format video images are obtained from the front to the back. Finally, a total of 20,160 complex scene multivehicle image sets T are extracted from the video at intervals of 10 seconds. Since the data acquisition takes full account of the problem that the positive sample number encountered by most datasets is zero, the design monitoring installation location is on each exit section of the loop. As shown in Figure 2, the number of image captures for the same vehicle is 2 times regardless of the intersection of any intersection from a to d (except for vehicles that repeatedly enter the road segment). A total of 45,742 identifiable vehicles with pixels greater than 128 are extracted from T and denoted as D. 80% of them are randomly selected to generate D_{train} of the training set and 20% to generate D_{test} of the training set.

5.2. Experimental Setup and Evaluation Criteria. Deep learning framework CAFFE [23] is used to implement the proposed method. The hardware configuration used in the experiment is as follows: GTX 1080 GPU, 8 GB video
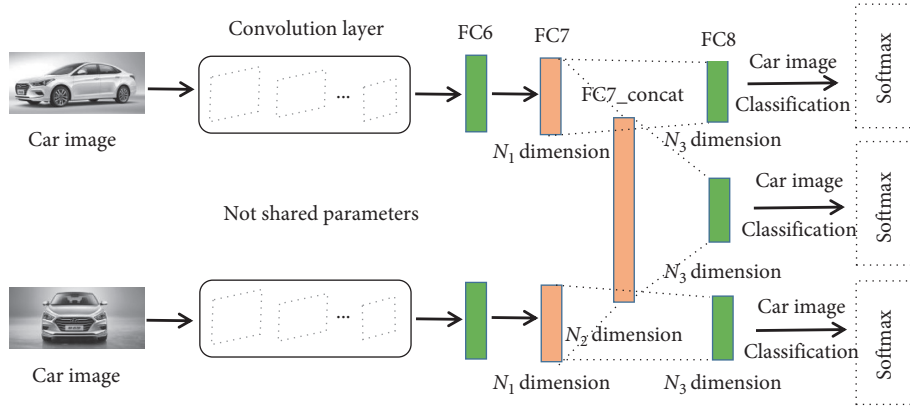
FIGURE 1: Illustration of double-channel symmetry CNN architecture.
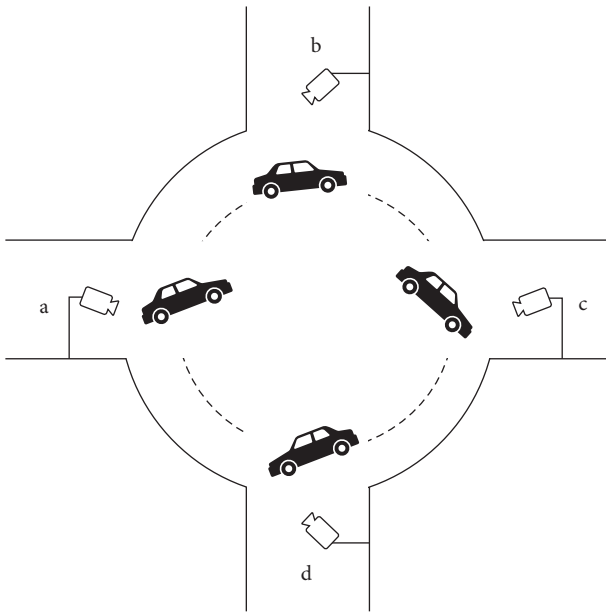

FIGURE 2: The installation location of 4 monitors.

memory, 128 GB memory, Intel core 8-core i7 processor CPU, and main frequency 3.60 GHZ.

Cumulative matching characteristic (CMC) curve, rank-1 accuracy, and mean average precision (MAP) were selected to evaluate the performance of the vehicle reidentification method. The CMC curve represents the probability that the truth value image to be queried will appear in a candidate sequence of different lengths. The rank-1 recognition accuracy rate represents the probability that the queried truth value image appears at the first position of the candidate sequence. The MAP is the average area under the curve of the accuracy rate of all query samples and recall rate, which reflects the overall performance of the vehicle reidentification method.

5.3. Experimental Result. By using the AlexNet and ResNet-50 model framework, the experimental results of the single-channel CNN method and the double-channel symmetric CNN method are compared, as shown in Table 1.

TABLE 1: The comparison between single-channel and double-channel symmetric.

	AlexNet		ResNet-50	
	Rank-1	MAP	Rank-1	MAP
Single-channel	57.14	33.52	73.65	47.11
Double-channel symmetric	62.27	37.78	74.36	49.55

The results show that the double-channel symmetric method has a stable improvement over the single-channel method. In the AlexNet model, the accuracy of rank-1 increased by 5.13%, and the accuracy of MAP increased by 4.26%. In the ResNet-50 model, the accuracy of rank-1 was improved by 0.71%, and the accuracy of MAP was improved by 2.44%.

In addition, on the ResNet-50 model, the accuracy of rank-1 and MAP of the proposed method in this study is 74.36% and 49.55%, respectively. At this time, the performance of this proposed vehicle recognition has reached a higher level.

The proposed method in this study is compared with some existing vehicle reidentification methods, including the traditional manual design method and the deep learning-based method. The specific comparison ends are shown in Table 2. The results show that the proposed method in this study has achieved a competitive performance of vehicle reidentification, which is better than some existing vehicle reidentification methods.

In order to further verify the effectiveness of this algorithm, the existing VeRi-776 dataset [28] is used for validation. The VeRi-776 dataset was captured by 20 cameras over a 24-hour urban area, containing 49357 images of 776 vehicles. Images are captured in a real-world unconstrained monitoring scenario and tagged with different attributes, such as type, color, and brand. Each vehicle is photographed by 2–18 cameras at different points of view, lighting, resolution, and occlusion. In the experiment of this study, each image of 2 cameras was selected as experimental data. The results are shown in Table 3. According to the data, the method proposed in this study achieves a better vehicle rerecognition performance, which is superior to other algorithms.

TABLE 2: The comparison results with self-built datasets.

Methods	Rank-1	MAP
SSDAL [24]	39.48	31.56
TMA [25]	42.38	36.97
End-to-end CAN [26]	54.82	37.59
SVDNet (AlexNet) [27]	51.93	35.52
SVDNet (ResNet-50) [27]	53.08	36.41
He et al. [28]	61.13	36.82
Proposed (AlexNet)	62.27	37.78
Proposed (ResNet-50)	74.36	49.55

TABLE 3: The comparison results with VeRi-776 datasets.

Methods	Rank-1	MAP
SSDAL [24]	40.59	32.31
TMA [25]	41.27	37.62
End-to-end CAN [26]	53.69	38.03
SVDNet (AlexNet) [27]	52.91	34.95
SVDNet (ResNet-50) [27]	54.24	37.32
He et al. [28]	62.48	37.46
Proposed (AlexNet)	63.31	37.98
Proposed (ResNet-50)	72.42	48.62

6. Conclusion

In order to further improve the performance of vehicle reidentification, this study proposes a double-channel symmetric CNN structure vehicle reidentification method. Under the original training samples, this algorithm inputs two samples at the same time, among which each sample has complementary characteristics. At this point, with limited training samples, the combination of inputs will be more diversified, which will enrich the training process of the CNN model. Therefore, the CNN model can be trained more fully, and a deeper learning model with stronger recognition ability can be obtained. The vehicle training map library was extracted from the monitoring video of different intersections, and then, the algorithm in this study was compared with other algorithms. Experimental results show that the vehicle recognition accuracy of the proposed algorithm is higher than other existing algorithms, which verifies the effectiveness of the proposed method.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] C. Zhang, L. Deng, Q. Du, and W. Deng, "Expressway vehicle management system based on vehicle face recognition," in *Proceedings of the International Conference on Man-Machine-Environment System Engineering*, pp. 369–376, Springer, Singapore, 2019.

[2] H. Chen and C. He, "A vehicle recognition algorithm based on fusion feature and improved binary normalized gradient feature," *Journal of Computational Methods in Sciences and Engineering*, vol. 19, no. 11, pp. 789–797, 2019.

[3] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *Journal of Big Data*, vol. 6, no. 1, p. 48, 2019.

[4] G. Manogaran, S. Baskar, P. M. Shakeel, N. Chilamkurti, and R. Kumar, "Analytics in real time surveillance video using two-bit transform accelerative regressive frame check," *Multimedia Tools and Applications*, vol. 79, pp. 16155–16172, 2020.

[5] Y. Chen and Z. H. Huo, "Person re-identification based on multi-directional saliency metric learning," *Journal of Image and Graphics*, vol. 20, no. 12, pp. 1674–1683, 2015.

[6] M. B. Qi, L. F. Hu, J. G. Jiang et al., "Person re-identification based on multi-features fusion and independent metric learning," *Journal of Image and Graphics*, vol. 21, no. 11, pp. 1464–1472, 2016.

[7] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition," *Computer Vision & Image Understanding*, vol. 150, pp. 109–125, 2016.

[8] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: a benchmark," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, IEEE, Santiago, Chile, December 2015.

[9] T. Berg and P. N. Belhumeur, "POOF: part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 955–962, IEEE, Portland, Oregon, June 2013.

[10] S. Kasamwattanarote, Y. Uchida, and S. I. Satoh, "Query bootstrapping: a visual mining based query expansion," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 2, pp. 454–466, 2016.

[11] Y. Zhou and L. Shao, "Vehicle re-identification by adversarial bi-directional lstm network," in *Proceedinds of the 2018 IEEE Winter Conference on Applications of Computer Vision*, Lake Tahoe, NV, USA, March 2018.

[12] H. Liu, Y. Tian, Y. Yang, P. Lu, and T. Huang, "Deep relative distance learning: tell the dif- ference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[13] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[14] 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 *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[15] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 2, Salt Lake City, UT, USA, June 2018.

[16] Y. Tang, D. Wu, Z. Jin, W. Zou, and X. Li, "Multi-modalmetric learning for vehicle re-identification in traffic surveillance environment," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2254–2258, Beijing, China, September 2017.

[17] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2018.

[18] F. Wu, S. Yan, J. S. Smith, and B. Zhang, "Joint semi-supervised learning and re-ranking for vehicle re-identification," in *Proceedings of the IEEE Conference on Pattern Recognition (ICPR)*, Beijing, China, August 2018.

[19] Y. Zhang, D. Liu, and Z.-J. Zha, "Improving triplet-wise training of convolutional neural network for vehicle re-identification," in *Proceedings of the IEEE International Conference on Multi- Media and Expo (ICME)*, pp. 1386–1391, Hong Kong, China, July 2017.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with deep Convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105, Curran Associates Inc., Lake Tahoe, Nevada, USA, December 2012.

[21] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE, Las Vegas, NV, USA, June 2016.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. F. Fei, "Imagenet: a large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, Miami, FL, USA, June 2009.

[23] Y. Q. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, ACM, Orlando, FL, USA, November 2014.

[24] C. Su, S. L. Zhang, J. L. Xing et al., "Deep attributes driven multi-camera person re-identification," in *Proceedings of the 14th European Conference on Computer Vision*, pp. 475–491, Springer, Amsterdam, Netherlands, October 2016.

[25] N. Martinel, A. Das, C. Micheloni et al., "Temporal model adaptation for person re-identification," in *Proceedings of the 14th European Conference on Computer Vision*, pp. 858–877, Springer, Amsterdam, Netherlands, October 2016.

[26] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[27] Y. F. Sun, L. Zheng, W. J. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3820–3828, IEEE, Venice, Italy, October 2017.

[28] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3997–4005, Long Beach, CA, USA, June 2019.