DUPL-VR: Deep Unsupervised Progressive Learning for Vehicle Re-Identification

Raja Muhammad Saad Bashir[1], Muhammad Shahzad[1], and Muhammad Moazam Fraz[1,2]

[1] School of Electrical Engineering and Computer Sciences (SEECS)
National University Sciences and Technology (NUST) Islamabad, Pakistan
{rbashir.mscs16seecs,muhammad.shehzad,moazam.fraz}@seecs.edu.pk
[2] The Alan Turing Institute, London, United Kingdom

Abstract. Vehicle re-identification (Re-ID) is a search for the similar vehicles in a multi-camera network usually having non-overlapping field-of-views. Supervised approaches have been used mostly for re-ID problem but they have certain limitations when it comes to real life scenarios. To cope with these limitations unsupervised learning techniques can be used. Unsupervised techniques have been successfully applied in the field of person re-identification. Having this in mind, this paper presents an unsupervised approach to solve the vehicle re-ID problem by training a base network architecture with a self-paced progressive unsupervised learning architecture which has not been applied to solve the vehicle re-ID problem. The algorithm has been extensively analyzed over two large available benchmark datasets VeRi and VehicleID for vehicle re-ID with image-to-image and cross-camera search strategies and the approach achieved better performance in most of the standard evaluation metrics when compared with the existing state-of-the-art supervised approaches.

1 Introduction

Vehicle re-identification (re-ID) plays an important role in an automated visual surveillance system [15]. Moreover, it also plays an important part in the monitoring or tracking the vehicles from multiple cameras in real time video surveillance or doing forensic analysis on the backup data for various kind of tasks e.g. patterns recognition of different vehicles, traffic conditions [20] etc. Vehicle re-ID is an automated process to find the similar vehicles in a multi-camera network normally having non-overlapping camera views. It is done using the unique ID's assigned to the vehicles upon discovery in the multi-camera network and keeping track of the discovered vehicles through the multi-camera network as depicted in Fig 1.

A trivial and easy to implement solution for vehicle re-ID is to match the license plates [21] as license plate number is a unique ID of the vehicle. But there are few issues to this approach e.g., low resolution, environmental factors (e.g. fog, dust, rain, storm etc.), side viewpoints (i.e., neither frontal nor backward) and poor/improper illumination. A solution to these issues is to use additional

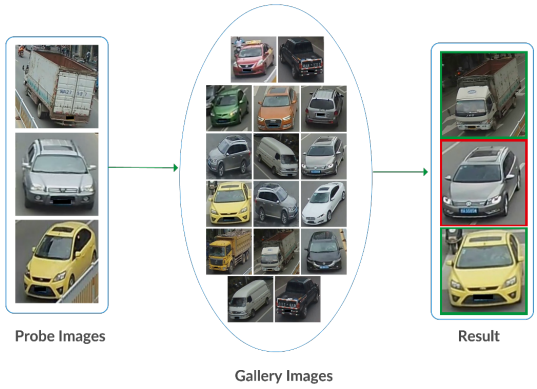

Fig. 1. Given the probe/query images finding the match in the gallery using similarity matching to find the most similar vehicles in gallery captured from different cameras having different viewpoints and illuminations.

appearance based approaches, which typically rely on visual and structural cues of the vehicles e.g., shape, color, texture [15][5] etc. However, these approaches work well in normal cases but are limited in terms of accuracy. With new techniques and advancements in neural network architectures especially Convolutional Neural Networks (CNNs) for computer vision tasks, the deep learning architectures have achieved the high accuracy and low error rate than the previous techniques and have enabled us to perform good in the real world scenarios. In this regard, This paper presents a semi-unsupervised approach to solve the vehicle re-ID problem by training a deep network architecture with a self-paced progressive unsupervised learning technique [8] enabling transfer of deeply learned representation towards unlabeled dataset. Following are the main contributions of the proposed approach.

- Exploiting a trained deep neural network architecture together with the self-paced learning scheme adopting the unsupervised K-means clustering that allows to infer the vehicles IDs in a semi-unsupervised manner.
- The analysis of the developed algorithm has been extensively analyzed over two large available benchmark datasets VeRi [12] and VehicleID [11] for vehicle re-ID with image-to-image and cross-camera searches.

2 Related Work

Due to numerous potential applications e.g., in surveillance and security, the vast amount of literature has framed the problem of re-ID in the domain of person re-ID. In this context, majority of the approaches solving the re-ID problem rely on conventional machine learning techniques [10, 17], Weighted Histogram Of Overlapping Stripes (WHOS) [16, 17], Bag of Words (BoW) [25] etc. These

features have been used to train traditional classifiers like Support Vector Machines (SVM) [1], AdaBoost [18] etc.
Vehicle re-ID is a closely related problem to person re-ID and the techniques developed in person re-ID may not directly applicable for vehicle re-ID domain as most of the feature descriptors (e.g., GOG, LOMO etc.) are specifically designed feature descriptors to re-identify persons and may not be able to handle monotone appearance of vehicles. Among the few existing works, Cheng-Hao *et al.*[9] presented the technique for detection and sub-categorization of vehicles in multi-view environments where they used a locally linear embedding and subsequently applied group sampling on the reduced data using k-means and further used the grouped data for training a boosted cascading tree classifier. Feris *et al.*[4] also proposed another technique for multi-view vehicle detection using the motion and shape-based features for training an AdaBoost classifier. Moreover, to overcome the occlusion problem, they adopted the Poisson distribution to synthetically generate the training data. Zapletal *et al.*[23] also presented a 3D bounding box based approach where they employed the color histogram and the histogram oriented gradients (HOG) features with linear regression to perform vehicle re-ID. In the context, of deep learning, Yang *et al.*[22] proposed a dataset COMPCARS for vehicle classification, where they applied the Convolutional Neural Network (CNN) [6] for feature extraction and later trained traditional learning classifiers for classification and attribute predictions. Liu *et al.*[12] also proposed a vehicle re-ID dataset (VeRi-776) and evaluated it using appearance based hybrid model comprising of both low and high level color/texture features extracted through handcrafted (SIFT, Bag-of-words) and CNN models to incorporate vehicle semantics. Liu *et al.*[14] later on proposed Progressive Vehicle re-ID (PROVID) model in which they divided the vehicle re-ID task into coarse-to fine search where the coarse filtering was performed using appearance based features and subsequently utilized Siamese Neural Network (SNN) [2] to match the license plates for accurate searching and later the vehicle re-ID was performed using spatio-temporal near-to-distant search. Liu *et al.*[11] also proposed a Deep Relative Distance Learning (DRDL) technique which uses the two branch CNN with coupled cluster loss for projecting raw vehicle features into Euclidian space and then measured the distance between different vehicles for similarity matching. Y. Zhang *et al.*[24] improved a similarity loss function called triplet method by introducing classification oriented loss triplet sampling of pair wise images to avoid misleading problems. Shen *et al.*[19] proposed a two stage technique where in first stage they used chain Markov random fields (MRF) model to create the spatio-temporal paths for each vehicle and then a Siamese-CNN and Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) called Path-LSTM model which takes the spatio-temporal path candidates and pair wise queries to perform vehicle re-ID using similarity scores. Liu *et al.*[13] improved their previous work by improving the first stage of coarse-to-fine search where they introduced the metric learning based approach Null-space-based FACT (NuFACT) where they project the multi-level features into a null space where they can use the Euclidean distance for finding the distance

between the vehicles.
As supervised algorithms have certain limitations that they require large amount of annotated data for training and limited to the dynamic growth of the data etc. So, using the unsupervised learning we can cope with these issues and in this context, inspired from the unsupervised techniques mentioned above, we have introduced these into vehicle re-ID We have proposed an approach that adopts a person re-ID technique [3] to solve the vehicle re-ID in an unsupervised manner.

3 Methodology

Figure 2 shows the block level diagram of the proposed system architecture highlighting the work flow of the input vehicles to the final output. It essentially formulates the whole vehicle re-ID problem into an unsupervised learning paradigm using a progressive two step approach explained below.

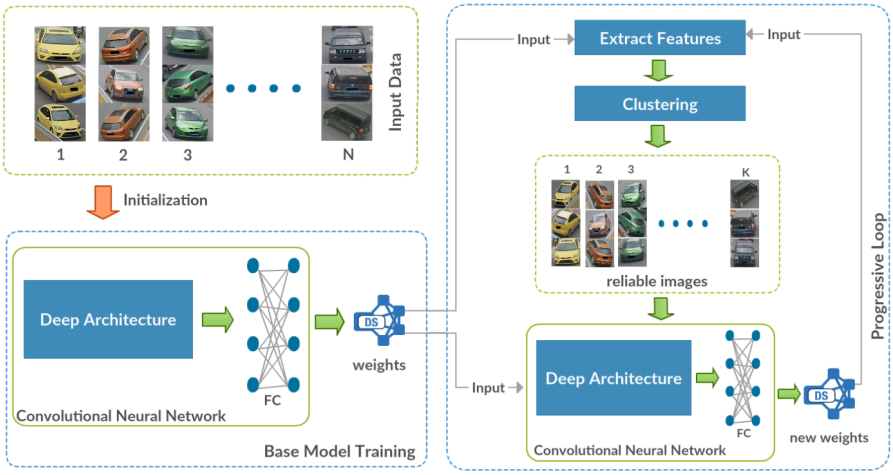

Fig. 2. The framework model for our vehicle re-ID. Irrelevant labeled data is fed into convolutional neural network for training and then trained model is stored in the first phase. Using the saved model features are extracted, clustered and used to find reliable images iteratively, which are then used to fine the base model.

3.1 Base Model Training

Lets suppose that a labeled dataset is given for training, the idea now is to replace the existing labels arbitrarily with unique numbers e.g. sequence number to allow latent space representation of the original labels denoted as $\{y_i\}_{i=1}^{M}$ with M being the total number of vehicles corresponding to $\{x_i\}_{i=1}^{N}$ where N is the

total number of images in the given dataset. Pre-processing using the data augmentation is applied and the data is fed to a deep CNN model (e.g., ResNet50 [7]) for training. The categorical cross entropy loss is used with stochastic gradient descent optimizer. The model is fine tuned till convergence i.e. the loss is not decreasing further and is almost constant.

3.2 Progressive Model Training

The base model is used to initialize and fine tune the progressive training part. This part includes the feature extraction, clustering and reliable feature selection as explained in the following subsections.

Feature Extraction and Clustering The features are obtained by removing the last classification layer from the deep model and using the average pooling layer as output i.e.

$$\mathbf{f}_i = \varphi(x_i, \theta) \text{ for all } i = 1, ..., N \tag{1}$$

where x_i the denotes the ith input image, θ are the weights and the φ represents the learned model which outputs the feature vector $\mathbf{f}_i$. The Extracted features $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_N\}$ are then fed to K-means algorithm to obtain a set $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k\}$ of k (= M) cluster centroids by minimizing the following optimization function:

$$\mathbf{C} \leftarrow \arg\min_{\mathbf{c}} \sum_{i=1}^{N} \sum_{j=1}^{M} \|\mathbf{f}_i - \mathbf{c}_j\|^2 \tag{2}$$

where $\mathbf{c}_{j=1,...,M}$ denotes the obtained M clusters centroid. To improve clusters further and avoid bad local minima due to some wrong assignments, we filtered the results by finding the dot product of features from their centroids. Features having distance more than a threshold lambda $\lambda = 0.85$ are filtered.

Model training and optimization The stable and refined cluster images are then used to fine tune the CNN model by using the centroids as labels $\mathbf{y}_i$ to formulate the training set. The model employed for training is ResNet50 [7] with some modifications i.e. an additional dropout layer tougher with a fully connected layer using SoftMax activation are added to the model. The model is initialized with the base model weights θ. The loss function for optimization is categorical cross entropy loss function L with stochastic gradient descent as an optimizer having the learning rate and momentum of 0.001 and 0.9 respectively. The model is thus fine-tuned on the obtained reliable training set of images. The process is iteratively performed where in each iteration, the training sample set is populated with increasingly robust/refined clusters enabling self-progressive and unsupervised learning until convergence i.e. we reach a minima and there is no more improvement. Table 1 depicts the whole algorithm used for training the proposed network architecture.

Algorithm 1 Unsupervised Vehicle Re-Identification

Input irrelavant labeled data $\{x_i\}_{i=1}^{N}$
No. of clusters K
Base model $\varphi(x, \theta_b)$
Output Model $\varphi(x, \theta_t)$

initialize $\theta_o \rightarrow \theta_b$;
while not convergence **do**
  extract features $\mathbf{f}_i = \varphi(x_i, \theta)$ for all $i = 1, ..., N$
  k-means and select centers $\mathbf{C} \leftarrow \arg\min_{\mathbf{c}} \sum_{i=1}^{N} \sum_{j=1}^{M} \|\mathbf{f}_i - \mathbf{c}_j\|^2$
  dot product $\frac{\mathbf{f}_i}{\|\mathbf{f}_i\|} \bullet \frac{\mathbf{c}_j}{\|\mathbf{c}_j\|} > \lambda$
  optimize the $\min_{\theta, \mathbf{w}} \sum_{i=1}^{r} L((\mathbf{y}_i, \varphi(\mathbf{x}_i, \theta)), \mathbf{w})$ function with reliable samples
  $\rightarrow \theta_t$
end while

4 Experimental Evaluation

There arent many notable architectures that solves the vehicle re-ID problem using unsupervised deep learning based techniques. However, to analyze and validate the performance of the proposed network architecture, we evaluated and compared the achieved results using state-of-the-art supervised deep learning based vehicle re-ID methods including PROgressive Vehicle re-ID (PROVID) and Deep Relative Deep Learning (DRDL). Before presenting the results, let's just analyze the datasets used.

4.1 Datasets

To validate the performance of the proposed approach, two different datasets VeRi [12] and VehicleID [11] have been used. VeRi is specially designed for the vehicle re-ID purpose having the train set of 576 vehicles and having 37781 images for vehicles and the test set having 200 vehicles and 11579 images with model and color information. VehicleID is a huge dataset as compared to the VeRi in terms of total images and vehicles, it has over 200,000 images and about 26,000 vehicles. Additionally, for about 9000 images and 10319 vehicles the model and color information are also present. The dataset is also divided into the train and test lists where train list has 13164 vehicles and it had multiple test sets based on the difficulty levels.

4.2 Performance Analysis

The performance of the given architecture has been validated using different test sets. The VehicleID dataset does not contain any query set so a query set is created from the test set by random selection from tests sets such that at least two images per vehicle are captured in the query set. In the following subsections,

the evaluation techniques metrics of performance of the proposed approach are discussed.

Evaluation Strategy and Metrics The trained architecture is evaluated on two strategies i.e. cross-camera search strategy where we have multi camera and vehicles information (VeRi dataset only) and image-to-image search strategy where we do not have multi-camera information (i.e., only vehicles information is given). In the first, the search query vehicles are not searched in the same camera but across the other cameras while in the latter, the search for query vehicles are performed in all the images regardless of the camera information. With these two strategies, The CMC curve, Rank@1, Rank@5 (precision at rank 1 and 5) are also used to find the accuracy of the methods along with mean average precision (mAP) to evaluate the comprehensive performance.

Table 1. image-to-image and cross-camera on VehicleID and VeRi

		image-to-image		**cross-camera**		
		Rank@1	**Rank@5**	**Rank@1**	**Rank@5**	**mAP**
	DRDL	45.41	63.70	-	-	-
VehicleID	NuFACT	43.72	65.19	-	-	-
	DUPL-VR	**71.45**	**81.69**	-	-	-
	PROVID	-	-	81.56	**95.11**	**52.42**
VeRi	DUPL-VR	100	100	**83.19**	91.12	40.05

Quantitative and Qualitative Results Table 4.2 provides the quantitative results obtained over the two employed datasets and the evaluation metrics. As can be seen that the proposed method outperforms the state-of-the-art supervised deep learning DRDL and PROVID methods in Rank@1 accuracy in both image-to-image and cross-camera scenarios. In Rank@5, the algorithm also demonstrates superior performance in image-to-image scenario and also shows competitive results in cross-camera scenarios. Figure 3 represents the CMC curve of the image-to-image search evaluation of the proposed DUPL-VR algorithm using VeRi and VehicleID. The orange line depicts the extra-ordinary performance of the proposed architecture over VeRi dataset. It is due to the reason that from same viewpoint multiple shots of the same vehicle are available at temporally close scales. Figure 4 depicts the CMC curve obtained in a cross-camera search scenario. As can be seen that the curve reaches 90% at Rank@5 meaning which shows that the proposed algorithm finds the correct results in 5 of the best matches.

Figure 4 show the CMC curve in blue shows the DUPL-VR on VeRi for different ranks from 1-50 where the Rank@1 represents the match is found in the first match and so on. VeRi performs quite well as we can see that the

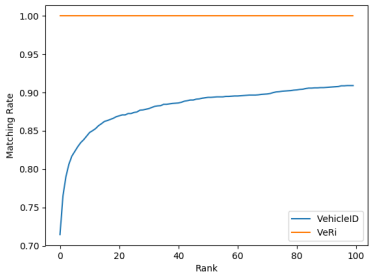

Fig. 3. CMC curve of VehicleID and VeRi in image-to-image strategy.

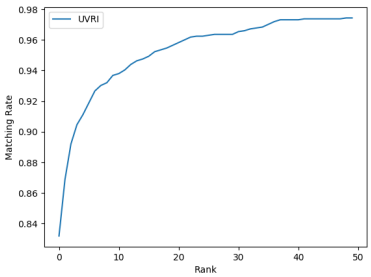

Fig. 4. CMC curve for the cross-camera search on VeRi only.

Fig. 5. Top 5 results of image-to-image search.

Fig. 6. Top 5 results of cross-camera search

curve reaches 90% of the corrects results threshold in the Rank@5 which tell us the high performance of the algorithm. Figure 5 and Figure 6 shows the obtained qualitative results over VehicleID and VeRi datasets using image-to-image and cross-camera search respectively.In Figure 5 (first row) the results are quite impressing, and the same car has been found in the top-5 and same is the case with the row 2-5. In 6th row we can see that the top-1 is quite same but is not the same but in top-4 and top-5 same van has been found. In the last row the cars found in top 3 places are the same. In Figure 6 (first row), we can see that the top-1 image is correctly being recognized while the top-2, top-3 and top-4 results have the same viewpoint but different color. Similarly, in the 2nd row, we can see that retrieved results have strong background appearance similarity (e.g., see the green bushes in top-1, top-2, top-3 and top-5). In third query we can see other cabs but due to same appearance that are hard to distinguish moreover in fourth same bus in appearance but with different license number plate can be seen as well and so on in other queries the results are based on appearance.

5 Conclusion

In this paper we have presented a semi-unsupervised deep learning-based vehicle re-ID technique. This technique uses the deep features extracted using the CNN and then uses the clustering and filtering to group and filter them for reliable selection of the input. Reliable clusters are then used to fine tuning of the model progressively until the convergence. This technique is evaluated on both datasets available VeRi and VehicleID for cross-camera and image-to-image search and it gave us the quite promising results then the supervised techniques. In cross-camera search it performs comparable to the supervised techniques while in image-to-image search it outperforms the supervised work. This technique has no limit on identities as you can train on certain number of identities and later use it on the test set regardless of the number of identities. In future work this technique can be further be improved by using improved CNN architectures better than ResNet50, using other improved clustering algorithms than k-means e.g. mean shift or some deep learning based unsupervised clustering algorithms. Reliable selection based on vehicle model and color will improve the result of reliable image selection and convergence. This does not ensure to optimally discriminate between different models of the vehicles having same color but certainly significantly help in overcoming the wrong and weak cluster assignment problems.

References

1. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recogn. Lett. **33**(7), 898–903 (May 2012)
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. pp. 737–744. NIPS'93, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
3. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. CoRR **abs/1705.10444** (2017)
4. Feris, R., Petterson, J., Siddiquie, B., Brown, L., Pankanti, S.: Large-scale vehicle detection in challenging urban surveillance environments. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV). pp. 527–533 (Jan 2011)
5. Feris, R.S., Siddiquie, B., Petterson, J., Zhai, Y., Datta, A., Brown, L.M., Pankanti, S.: Large-scale vehicle detection, indexing, and search in urban surveillance videos. IEEE Transactions on Multimedia **14**(1), 28–42 (Feb 2012)
6. Feris, R.S., Siddiquie, B., Petterson, J., Zhai, Y., Datta, A., Brown, L.M., Pankanti, S.: Large-scale vehicle detection, indexing, and search in urban surveillance videos. IEEE Transactions on Multimedia **14**(1), 28–42 (Feb 2012)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
8. Jiang, L., Meng, D., Yu, S.I., Lan, Z., Shan, S., Hauptmann, A.G.: Self-paced learning with diversity. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. pp. 2078–2086. NIPS'14 (2014)
9. Kuo, C.H., Nevatia, R.: Robust multi-view car detection using unsupervised sub-categorization. In: 2009 Workshop on Applications of Computer Vision (WACV). pp. 1–8 (Dec 2009)

10. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2197–2206 (June 2015)
11. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2167–2175 (June 2016)
12. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: 2016 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (July 2016)
13. Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia **20**(3), 645–658 (March 2018)
14. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 869–884. Springer International Publishing, Cham (2016)
15. Matei, B.C., Sawhney, H.S., Samarasekera, S.: Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In: CVPR 2011. pp. 3465–3472 (June 2011)
16. Mubariz, N., Mumtaz, S., Hamayun, M.M., Fraz, M.M.: Optimization of person reidentification through visual descriptors. In: Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,. pp. 348–355 (2018)
17. Mumtaz, S., Mubariz, N., Saleem, S., Fraz, M.M.: Weighted hybrid features for person re-identification. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6 (Nov 2017)
18. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference
19. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1918–1927 (Oct 2017)
20. Sivaraman, S., Trivedi, M.M.: Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems **14**(4), 1773–1795 (Dec 2013)
21. Watcharapinchai, N., Rujikietgumjorn, S.: Approximate license plate string matching for vehicle re-identification. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6 (Aug 2017)
22. Yang, L., Luo, P., Loy, C.C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3973–3981 (June 2015)
23. Zapletal, D., Herout, A.: Vehicle re-identification for automatic video traffic surveillance. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1568–1574 (June 2016)
24. Zhang, Y., Liu, D., Zha, Z.J.: Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 1386–1391 (July 2017)
25. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person reidentification: A benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1116–1124 (Dec 2015)