Large Scale Vehicle Re-Identification by Knowledge Transfer from Simulated Data and Temporal Attention

Viktor Eckstein[1,3] Arne Schumann[1,2] Andreas Specker[1,3]
[1]Fraunhofer IOSB, Karlsruhe, Germany [2]Fraunhofer Center for Machine Learning
[3]Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
eckstein1991@gmail.com {arne.schumann, andreas.specker}@iosb.fraunhofer.de

Abstract

Automated re-identification (re-id) of vehicles is the foundation of many traffic analysis applications across camera networks, e.g. vehicle tracking, counting, or traffic density and flow estimation. The re-id task is made difficult by variations in lighting, viewpoint, image quality and similar vehicle models and colors that can occur across the network. These influences can cause a high visual appearance variation for the same vehicle while different vehicles may appear near identical under similar conditions. However, with a growing number of available datasets and well crafted deep learning models, much progress has been made. In this work we summarize the results of our participation in the NVIDIA AI City Challenge 2020 for vehicle re-id. We address the re-id task by relying on well-proven design choices from the closely related person re-id literature. In addition to this, we explicitly address viewpoint and occlusions variation. The former is addressed by incorporating vehicle viewpoint classification results into our matching distance. The required viewpoint classifier is trained predominantly on simulated data and we show that it can be applied to real-world imagery with minimal domain adaptation. We address occlusion by relying on temporal attention scores which emphasize video frames that contain minimal occlusion. Finally, we further boost re-id accuracy by applying video-based re-ranking and an ensemble of complementary models. Our models, code, and simulated data is available at https://github.com/corner100/2020-aicitychallenge-IOSB-VeRi.

1. Introduction

The growing numbers of vehicles in our streets make detailed traffic analysis and statistics an important foundation for city and infrastructure planning and management. Automated analysis of traffic flow can be conducted based on the large amounts of available traffic cameras. Besides vehicle detection and tracking, the re-identification (re-id) of vehicles is a key task in this context. Vehicle re-id facilitates cross-camera vehicle tracking, cross-camera vehicle counting and automated vehicle search. The problem of vehicle re-id is often posed as a retrieval task, where a query image is given and further instances of the depicted vehicle must be found in a gallery of images. The result is a list of gallery images ranked by similarity to the query. In the last few years, growing attention is payed to the task of vehicle re-id, due to newly available datasets, e.g., [13], and much progress in the related field of person re-id [17]. Re-id approaches must address a number of challenges which influence the visual appearance of vehicles, see Figure 1. These challenges include variation in lighting conditions, different viewpoints across cameras, occlusions by other vehicles or roadside elements, a range of image quality due to camera distance and heterogeneous camera networks. These influences on visual appearance can lead to a large visual variation of the same vehicle (i.e. high inter class variation) while the limited range of vehicle models and colors can cause different vehicles to look quite similar (i.e. low intra class variation). Typically, the re-id task is addressed by learning or designing a feature space or distance measure, which reduces inter class variation and increases intra class

Figure 1. Example images and challenges from the CityFlow dataset [13]. The first column shows the same vehicle under different lighting and viewpoint. In the second column strong occlusion is illustrated, as well as extreme viewpoints. Column three exhibits differences in image quality and column four shows different vehicles with high visual similarity.

variation to achieve better separability.

In this work, we propose a vehicle re-identification approach, which relies on a number of design aspects from the extensive person re-id literature to form a strong baseline model. We rely on these well established design choices to address the class variation problem but incorporate our own modifications to address specific challenges which we deem to impact re-id accuracy most strongly, namely viewpoint and occlusions. To address viewpoint, we expand on the baseline model by incorporating orientation information learned on simulated vehicle data. By including this orientation information into our distance measure, we can notably improve accuracy. We rely on video data to alleviate the problem of temporary occlusions and avoid their negative impact by introduction of a temporal attention module. Finally, we apply video-based re-ranking to improve our retrieval results and form an ensemble of our strongest models to combine strengths and compensate for weaknesses of individual models. Our proposed approach achieved 4^{th} place in the NVIDIA AI City Challenge 2020.

2. Related Work

Our discussion of related work will be limited to approaches most directly related to our method and application. Specifically approaches relying on parsing license plates, e.g., [6], will not be discussed, as they are often not applicable due to low image quality.

Most current methods for vehicle re-identification are based on deep learning models, as these have proven very accurate for a number of computer vision tasks. In [20] a viewpoint-aware attentive multi-view inference model is proposed by Zhou et al. to solve the re-id problem. Wang et al. [14] extract features in local regions of different orientations based on 20 vehicle keypoints to encode the fine localized details of vehicles. In [10] Shen et al. propose a two-stage framework for vehicle re-id. The first stage generates a series of candidate spatio-temporal paths with the query images as the start and end state. Then a siamese CNN and LSTM is used to better leverage the spatio-temporal regularization from the candidate path. Zhu et al. [21] propose quadruple directional deep learning networks to obtain improved viewpoint robust deep learning features of vehicle images. Huang et al. [4] apply a similar attention module to ours in order to identify the most relevant frames in a vehicle track.

2.1. Simulated Data

Synthesizing image data is a cost-effective way to obtain highly accurate labeled data in large amounts. This comes at the cost of a reality gap which exists between simulated and real-world data. This gap often prevents models from directly applying to real world data and requires additional domain adaptation. In the context of vehicle re-id, Tang et al. [12] have used simulated vehicle data with annotations for color, id, and pose information to train a multi task re-id model. Yao et al. [15] introduce a vehicle rendering engine and apply it to generate simulated data, which is adapted to target real-world data by adapting a set of attributes. In contrast to both these methods, which include simulated data directly into the learning process, we use the simulated data to train an orientation classifier and transfer this orientation information to our real world data.

2.2. Re-Ranking

Re-ranking is an often applied post-processing step to re-id, which can improve ranking results. Methods typically rely on the consistency and nearest-neighbor relationships of gallery images based on the initial re-id ranking. Re-ranking approaches are widely used in person re-identification [16, 19, 9]. One of the most widely used methods by Zhong et al. [19] relies on computation of the Jaccard distance based on k-reciprocal neighbor lists defined by the original ranking. Re-ranking methods do not require additional training and are thus frequently used to improve accucacy.

3. Methodology

We used the strong baseline framework of Luo et al. [7]. While the strong baseline model was developed for person re-identification, all aspects are class agnostic and prove to achieve very strong results for vehicle re-id as well. Most notably, we apply softmax loss for classification and a combination of triplet and center loss to achieve improved vehicle separability in feature space. The losses are separated by batch normalization, following the observations made in [7]. We use learning rate warmup, random erasing data augmentation, label smoothing, and modification of stride as proposed in [7] but forego the center loss, as it did not improve resulting matching accuracies.

3.1. Synthetic Data for Vehicle Orientation

Variation in viewpoint between camera and vehicle has one of the largest impacts on visual appearance and matching result. Thus, an explicit inclusion of viewpoint information into the matching process harbors great potential to increase accuracy. However, orientation is not straightforward to parametrize and annotate. Coarse annotations, such as *front, side, back* are easily understood by human annotators but lack nuance for use in machine learning models and can often lead to inconsistent judgements by annotators. More detailed information, such as vehicle keypoints are very costly to annotate manually. Direct annotation of relative angles between camera and vehicle is very hard to accurately judge for annotators. Synthetic vehicle data on the other hand allows to accurately set these properties and generate correspondingly annotated and consistent data in

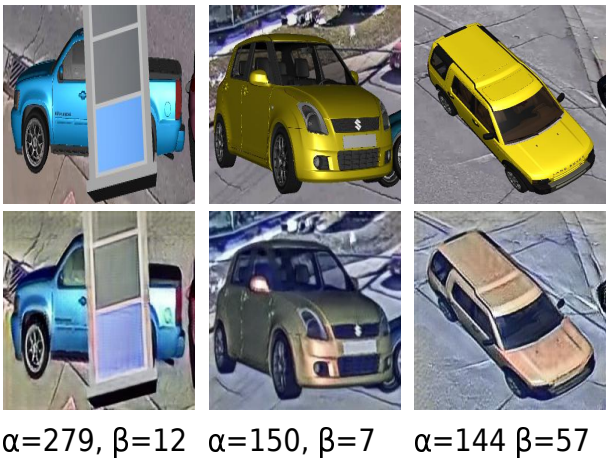

Figure 2. Example images of our generated synthetic dataset with camera position information. The top row shows images before domain adaptation, bottom row depicts the same images after adaptation with SPGAN [2].

large amounts. While direct use of synthetic imagery for the training of re-id models is possible, the reality gap between synthetic and real-world data has an impact on model accuracy and typically requires complex mechanisms to alleviate this detrimental effect. Rather than using images directly to leverage knowledge from the synthetic domain, it is our intuition that a more robust knowledge transfer can be achieved on a meta level. We thus aim to use the synthetic data to train a view regression model. View regression is a simpler problem than re-id as fine details and image characteristics should not strongly affect the result. Thus, transfer of information on this level should be more robust across the domain gap and minimal domain adaptation could suffice to use the view information on real-world data.

In order to generate extensive amounts of simulated vehicle data, we rely on the VehicleX engine [15]. We extended this tool to provide two angles with each image, as depicted in Figure 2. The in-plane rotation of the camera relative to the vehicle, i.e., $0°$ for back view, $180°$ for frontal, etc., is represented by α. The out-of-plane rotation, β, ranges from $0°$ at ground plane to $90°$ directly above the car. In order to prepare the values for training of a regression model, we transform the angles into positional values in x, y, z-coordinates with $x = sin(\alpha)$, $y = cos(\alpha)$ and $z = sin(\beta)$. x and y ranges from -1 to 1, from back to front and y from -1 to 1, left to right. z ranges from 0 to 1, from parallel to above the car. We generate a total of 191,116 images. Using this data, we perform domain adaptation to the CityFlow dataset by training a style transfer model, SPGAN [2], using the CityFlow train set.

We then train a CNN, the orientation model, to regress the three parameters x, y and z. As backbone model we choose ResNet-152 pretrained on ImageNet and modify the last layer for our three output values. x and y parameters are predicted with a $tanh$ activation function and z with a sigmoid activation function.

After training this orientation model on the synthetic data we labeled a subset of the real training data with the help of the pretrained orientation model. For the final solution we used the orientation model trained on the labeled dataset which is a subset of the real images from the training data.

We use the obtained orientation information to modify our matching distance. Since vehicles with similar angles are more likely to be close in the feature space we artificially increase matching distance in such cases. For a similar camera view τ_1 we add a constraint σ to punish cars from gallery images with similar camera views as the query image. Galleries that have a very close distance τ_2 to the query image are not punished in this manner, as they typically represent correct matches. See Equation 1 and Figure 4.

$$d_{new} = \begin{cases} d + \sigma & \text{if } d_{orient} < \tau_1 \ \& \ d > \tau_2 \\ d & \text{else} \end{cases} \quad (1)$$

3.2. Video-based Re-Id

For the main re-id task we first train a ResNet-152 baseline model for image-to-image re-id (I2I) with real and synthetic data. The weights of this model are then used as initialization for our video-to-video (V2V) model. For V2V we use only the real world CityFlow data, since no tracks are available for the synthetic data.

Temporal Attention The base model of the I2I network is used to generate feature vectors for every image in a vehicle track. The resulting feature vectors are then fed into a temporal attention module based on [4] to generate a single feature vector as representation of the track. For the temporal attention head, see Figure 3, of the V2V model we apply feature padding by repeating the first and last frame's feature vector $l/2$ times. We then apply a convolutional layer with the input and output size same as the feature vector. The kernel size is set to l, i.e., l images are considered for attention computation. Softmax is the applied to obtain results between 0 and 1 which are used for weighting. Notably, we do not just learn a single weight for each frame's feature but rather one weight for each feature dimension. Thus, our attention mechanism combines primarily focuses on the temporal aspect but can still maintain important specific information from individual feature dimensions in otherwise low-weighted frames. For training, video track length was set to a fixed size l. If there are more than l images in the track, l random consecutive images are chosen. If there are fewer images, the first and last images are repeated. During evaluation the batch size was chosen to be one and the original video track length was used.

Re-Ranking A simple but very effective post processing

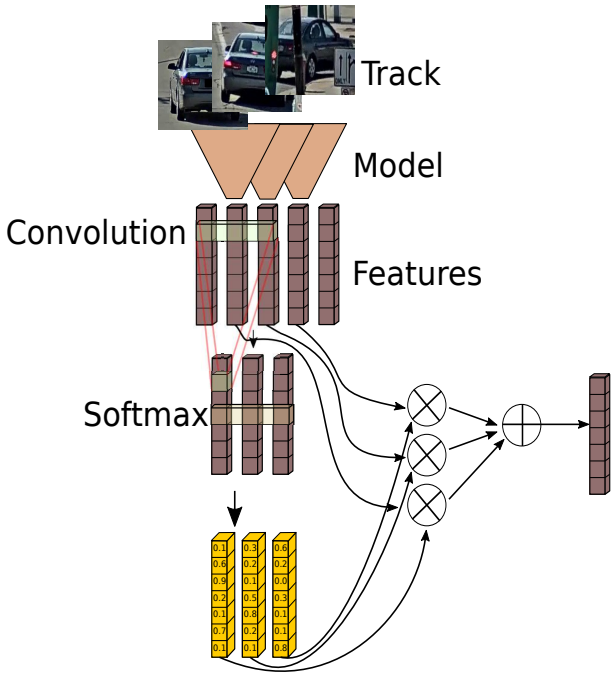

Figure 3. Our temporal attention head for the V2V model. We do not only weight every feature vector but also every feature.

step is V2V re-ranking [19]. Since the V2V model provides a feature vector per track we can use the same re-ranking method as for I2I.

Ensembling Model ensembling is a popular strategy, as it can eliminate weaknesses of individual models while combining their strengths. He et al. [3] show that different image sizes yield different performance. We show that ensembling such different models can significantly improve the score. We first create an ensemble of two different re-id models by adding their resulting distance matrices. Since both models have the same backbone architecture we do not have to weight them explicitly but rather compute a simple average.

4. Evaluation

We evaluate our proposed model on the CityFlow dataset [13], which is provided as part of the AI City Challenge. The dataset contains 666 distinct vehicles and 40 cameras. It is split into 36,935 images for training with 1897 tracks and 19,342 in evaluation with 1052 query images and 18290 gallery images comprising 798 tracks. The average track length in the train set is 19.47 with a minimum length of 1 and a maximum of 56. In the test set the average track length is 22.92 and ranges from 2 to 58 images per track. We apply mean Averaged Precision (mAP) as our evaluation metric. The mAP summarizes the accuracy of ranked lists across all queries in the test set in a single value.

For our final submission to the AI City Re-Id Challenge 2020, five models were trained. 1) A ResNet-152 with 256×256 input images as I2I re-id task. The pretrained weights where used for 2) the V2V ResNet-152 on tracks. 3) A ResNet-152 with 320×320 input images for I2I and again for 4) V2V and 5) a ResNet-152 for camera position estimation. The I2I models were trained on the real and synthetic data provided by the organizers with ImageNet pretrained weights. The I2I model was used for the V2V base model as the pretrained weights.

We trained our network for 80 epochs and saved the weights in every epoch. For the submission we choose weights that performed best on the validation set. The I2I models where trained on the same parameters as [7] with a learning rate of 0.0001, softmax and triplet loss and 10 warmup iterations. The V2V models where trained with a learning rate of $1.0e-9$. For V2V the track length l was 5. And we choose the re-ranking parameters to be $k_1 = 6, k_2 = 3$. The feature vectors where generated for the query and gallery tracks with both V2V-models. Three distance matrices were computed. One for each V2V model and one from the orientation model using the output of the last layer. The three distance matrices were combined described in 3. Parameters $\tau_1 = 0.03$, $\tau_2 = 0.3$ and $\sigma = 0.5$ were chosen.

In Table 2 we observe that bigger image size achieves better results. V2V re-id performs better that I2I and V2V re-ranking improves the mAP significantly. We thus choose

Table 1. Results for a ResNet-152 I2I model on the real dataset and an additional syntethic dataset. We could not achieve better results by generating our own synthetic dataset.

Synthetic Dataset	mAP
From organizers	**40.71**
Ours before domain adaptation	38.35
Ours after domain adaptation	37.20

Table 2. The mAP for the CityFlow-ReID Track2 dataset for different methods with re-ranking method (rr), image-to-image (I2I) and video-to-video (V2V).

Method	Im.size	rr	mAP
I2I	256×256		43.3
I2I	256×256	✓	49.4
I2I	320×320		45.5
I2I	320×320	✓	52.3
V2V	256×256		55.7
V2V	256×256	✓	62.3
V2V	320×320		57.3
V2V	320×320	✓	**63.7**

Figure 4. Ranking results before (top) and after (bottom) distance modulation with the orientation information. The blue image is the query, green the positive and red the negative galleries. Images with the same camera view are more likely to have a small distance, therefore we add a constant σ to punish them. Images with a very small distance are not punished, see Equation 1

Table 5. Final results for the AI City 2020 Challenge for vehicle re-id (track2).

Team	mAP	rank
Baidu-UTS	84.13	1
RuiYanAI	78.10	2
DMT	73.22	3
IOSB-VeRi	**68.99**	**4**
BestImage	66.84	5
BeBetter	66.83	6
UMD_RC	66.68	7

Table 3. The results for the AI City re-id task for our ensembe with re-ranking (rr) and video-to-video (V2V) and inclusion of orientation information.

Method	mAP
V2V 256×256 rr + V2V 320×320 rr	65.5
V2V 256×256 rr + V2V 320×320 rr + Orientation	**69.0**

Table 4. The results for the AI City re-id task 2019 compared with ours. Marked with $*$ is the best method in 2019 with no use of added real-world data from either external sources or other tracks of the 2019 challenge.

Team	External Data	Other Tracks	mAP
Baidu Zero One [11]	✓	✓	85.54
U. Washington IPL [4]	✓		79.17
Australian National U. [8]		✓	75.89
U. Tech. Sydney [18]	✓	(✓)	75.60
BUPT Traffic Brain [3]	✓		73.02
U. Maryland RC [5]	✓		60.78
INRIA STARS [1]			53.44*
Ours			68.99

the strongest V2V models with different resolutions for our ensemble. In Table 3 we can see that ensembling such models with different input resolutions is very effective. Furthermore, our proposed distance modification using the orientation information can improve the resulting accuracy of the ensemble notably.

Additionally, we compare our result with the published approaches of the 2019 AI City Challenge in Table 4. Many of these methods relied on use of external data or leveraging data from the separate city scale vehicle tracking task in the 2019 challenge. This practice was prohibited in the 2020 installment of the challenge and our proposed method outperforms all published 2019 approaches, which do not rely on such additional data.

In Table 5 the final results of the 2020 challenge are given. As details of individual approaches are not yet published, we can not discuss the specific differences between our and other methods. Our approach ranks fourth place out of 41 teams.

5. Conclusion

In summary, we have introduced a vehicle re-id model based on best practices from the larger person re-id literature. We extended the model by specifically addressing the challenges of viewpoint variation and occlusions through a viewpoint classifier trained on simulated data and a temporal attention mechanism to better leverage video data. Both modifications lead to a notable increase in matching accuracy, resulting in an overall mAP of 68.99% on the CityFlow dataset. Our final result placed fourth out of 41 teams in the 2020 AI City challenge in the re-id track.

References

[1] Hao Chen, Benoit Lagadec, and Francois Bremond. Partition and reunion: A two-branch neural network for vehicle re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[2] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, 2018. 3

[3] Zhiqun He, Yu Lei, Shuai Bai, and Wei Wu. Multi-camera vehicle tracking with powerful visual features and spatial-temporal cue. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 4, 5

[4] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification

using temporal attention model and metadata re-ranking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 3, 5

[5] Pirazh Khorramshahi, Neehar Peri, Amit Kumar, Anshul Shah, and Rama Chellappa. Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[6] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision*, pages 869–884. Springer, 2016. 2

[7] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 4

[8] Kai Lv, Heming Du, Yunzhong Hou, Weijian Deng, Hao Sheng, Jianbin Jiao, and Liang Zheng. Vehicle re-identification with location and time stamps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[9] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018. 2

[10] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang 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*, pages 1900–1909, 2017. 2

[11] Xiao Tan, Zhigang Wang, Minyue Jiang, Xipeng Yang, Jian Wang, Yuan Gao, Xiangbo Su, Xiaoqing Ye, Yuchen Yuan, Dongliang He, Shilei Wen, and Errui Ding. Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[12] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 211–220, 2019. 2

[13] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 1, 4

[14] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 379–387, 2017. 2

[15] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. *arXiv preprint arXiv:1912.08855*, 2019. 2, 3

[16] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016. 2

[17] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 1

[18] Zhedong Zheng, Tao Ruan, Yunchao Wei, and Yi Yang. Vehiclenet: Learning robust feature representation for vehicle re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5

[19] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 2, 4

[20] Yi Zhou and Ling Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6489–6498, 2018. 2

[21] Jianqing Zhu, Huanqiang Zeng, Jingchang Huang, Shengcai Liao, Zhen Lei, Canhui Cai, and Lixin Zheng. Vehicle re-identification using quadruple directional deep learning features. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 2