Vehicle Re-Identification in Context

Aytaç Kanacı[1], Xiatian Zhu[2], and Shaogang Gong[1]

[1] Queen Mary University of London, London E1 4NS, UK
[2] Vision Semantics Limited, London E1 4NS, UK
{a.kanaci,s.gong}@qmul.ac.uk
eddy@visionsemantics.com

Abstract. Existing vehicle re-identification (re-id) evaluation benchmarks consider strongly artificial test scenarios by assuming the availability of high quality images and fine-grained appearance at an almost constant image scale, reminiscent to images required for Automatic Number Plate Recognition, e.g. VeRi-776. Such assumptions are often invalid in realistic vehicle re-id scenarios where arbitrarily changing image resolutions (scales) are the norm. This makes the existing vehicle re-id benchmarks limited for testing the true performance of a re-id method. In this work, we introduce a more realistic and challenging vehicle re-id benchmark, called Vehicle Re-Identification in Context (VRIC). In contrast to existing vehicle re-id datasets, VRIC is uniquely characterised by vehicle images subject to more realistic and unconstrained variations in resolution (scale), motion blur, illumination, occlusion, and viewpoint. It contains 60,430 images of 5,622 vehicle identities captured by 60 different cameras at heterogeneous road traffic scenes in both day-time and night-time. Given the nature of this new benchmark, we further investigate a multi-scale matching approach to vehicle re-id by learning more discriminative feature representations from multi-resolution images. Extensive evaluations show that the proposed multi-scale method outperforms the state-of-the-art vehicle re-id methods on three benchmark datasets: VehicleID, VeRi-776, and VRIC[3].

1 Introduction

Vehicle re-identification (re-id) aims at searching vehicle instances across non-overlapping camera views by image matching [14]. Influenced by the recent extensive studies on person re-id [6,25,?,10,?,21,?,?,30], vehicle re-id has started to gain increasing attention in the past two years, which promises the potential for more flexible means for vehicle recognition and search than Automatic Number Plate Recognition (ANPR). However, vehicle re-id by visual appearance is a challenging task due to the very similar appearance of different vehicle instances of the same model type and colour, and a significant visual appearance variation of the same vehicle instance in different camera views.

Current vehicle re-id studies are mainly driven by two benchmark datasets, VehicleID [14] and VeRi-776 [16]. While having achieved significant performance

[3] Avaliable at http://qmul-vric.github.io

improvement (e.g. from 61.44% by [16] to 92.35% Rank-1 by [23] on VeRi-776), the scalability of existing re-id algorithms to real-world vehicle re-id applications remains unclear. This is because existing benchmarks represent somewhat rather artificial tests using high-quality images of high resolution, no motion blur, limited weather conditions and occlusion (Table 1 and Fig 1). This is more reminiscent to imaging conditions for ANPR than what is typical for vehicle re-id in wide-view traffic scenes "in-the-wild".

In this work, we introduce a new benchmark dataset called **Vehicle Re-Identification in Context** (VRIC) for more realistic and challenging vehicle re-identification. VRIC consists of 60,430 images of 5,656 vehicle IDs collected from 60 different cameras in traffic scenes. VRIC differs significantly from existing datasets in that *unconstrained* vehicle appearances were captured with variations in imaging resolution, motion blur, weather condition, and occlusion. This VRIC dataset aims to provide a more realistic vehicle re-id evaluation benchmark.

We make two contributions: (1) We create and introduce a more realistic vehicle re-id benchmark VRIC that contains vehicle images of *unconstrained* visual appearances with variations in resolution, motion blur, weather setting, and occlusion. This dataset is created from the UA-DETRAC benchmark [24] originally designed for object detection and multi-object tracking in traffic scenes, therefore reflecting appropriately and providing the necessary vehicle re-id environmental context and viewing conditions. This new benchmark will be publicly released. (2) We further investigate a Multi-Scale (resolution) Vehicle Feature (MSVF) learning model to address the inherent and significant multi-scale resolution in vehicle visual appearances from typical wide-view traffic scenes, currently an unaddressed problem in vehicle re-id due to the lack of a suitable benchmark dataset. Extensive comparative evaluations demonstrate the effectiveness of the proposed MSVF method in comparison to the state-of-the-art vehicle re-id techniques on the two existing benchmarks (VehicleID [14] and VeRi-776 [16]) and the newly introduced VRIC benchmark.

Table 1: Characteristics of vehicle re-id datasets.

Dataset	Images	IDs	Cameras	Resolutions Width×Height (Mean)	Motion Blur	Illumination	Occlusion
VehicleID [14]	113,123	15,524	-	345.4×376.1	No	Limited	No
VeRi-776 [16]	51,034	776	20	376.1×345.4	No	Limited	No
VD1 [26]	846,358	141,756	-	424.8×411.0	No	Limited	No
VD2 [26]	690,518	79,763	-	401.3×376.4	No	Limited	No
VRIC (Ours)	60,430	5,622	120	65.9×103.0	Unconstrained	Unconstrained	Unconstrained

2 Related Work

Vehicle Re-Identification. Whist vehicle re-id is less studied than person re-id [6,10,2,25,11,21,12,30,3,**?**,**?**], there are a handful of existing methods. Notably,

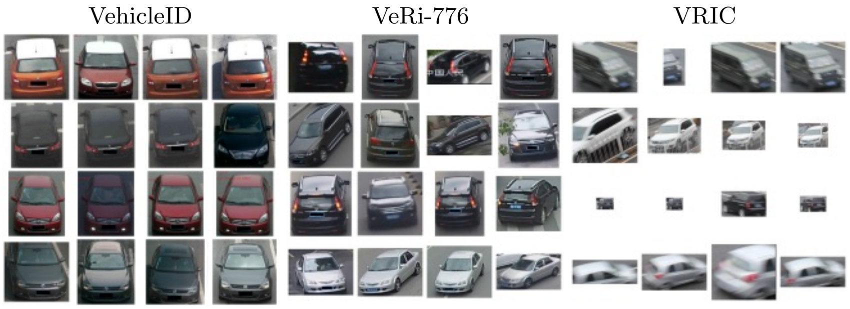

Fig. 1: Example images of VehicleID, VeRi-776 and VRIC. Images in each row depict the same vehicle instance. VRIC images exhibit significantly more unconstrained variations in resolution, motion blur, occlusion/truncation and illumination within each vehicle bounding-box images.

Feris *et al.*[5] proposed an attribute-based re-id method. The vehicles are firstly classified by different attributes like car model types and colours. The re-id matching is then conducted in the attribute space. Dominik *et al.*[28] used 3D bounding boxes for rectifying car images and then concatenate colour histogram features of vehicle image pairs. A binary linear SVM model is then trained to verify whether a pair of images have the same identity. Both methods rely heavily on weak hand-crafted visual features in a complex multi-step based approach, suffering from weak discriminative model generalisation.

More recently, deep learning techniques have been exploited to vehicle re-id. Liu *et al.*[16] explored a deep neural network to estimate the visual similarities between vehicle images. Liu *et al.*[14] designed a Coupled Clusters Loss (CCL) to boost a multi-branch CNN model for vehicle re-id. Kanaci [**?**] explored the appearance difference at the coarse-grained vehicle model level. All these methods utilise the global appearance features of vehicle images and ignore local discriminative regions. To explore local information and motivated by the idea of landmark alignment [29] in both face recognition [22] and human body pose estimation [18], Wang *et al.*[23] considered 20 vehicle keypoints for learning and aligning local regions of a vehicle for re-id. Clearly, this approach comes with extra cost of exhaustively labelling these keypoints in a large number of vehicle images, and the implicit assumption of having sufficient image resolution/details for computing these keypoints.

Additionally, space-time contextual knowledge has also been exploited for vehicle re-id subject to structured scenes [16,19]. Liu *et al.*[16] proposed a spatio-temporal affinity approach for quantifying every pair of images. Shen *et al.*[19] further incorporated spatio-temporal path information of vehicles. Whilst this method improves the re-id performance on the VeRi-776 dataset, it may not generalise to complex scene structures when the number of visual spatio-temporal path proposals is very large with only weak contextual knowledge available to facilitate model decision.

In contrast to all existing methods as above, we address a different problem of learning multi-scale feature representation for vehicle re-id.

Vehicle Re-Identification Benchmarks. There are in total four vehicle re-id benchmarks reported in the literature. Liu *et al.* [14] introduced the "VehicleID" benchmark with a total of 221,763 images from 26,267 IDs. In parallel, Liu *et al.* [15] created "VeRi-776", a smaller scale re-id dataset (51,035 images of 776 IDs) but with space-time annotations among 20 cameras in a road network. Recently, Yan *et al.*[26] presented two larger datasets (846,358 images of 141,756 IDs in "VD1", 690,518 images of 79,763 IDs in "VD2") with similar visual characteristics as VehicleID.

Whilst these existing benchmarks have contributed significantly to the development of vehicle re-id methods, they only represent *constrained* test scenarios due to the rather artificial assumption of having high quality images of constant resolution (Table 1). This makes them limited for testing the true robustness of re-id matching algorithms in typically *unconstrained* wide-view traffic scene imaging conditions. The VRIC benchmark introduced in this work addresses this limitation by providing a vehicle re-id dataset conditions giving rise to changes in resolution, motion blur, weather, illumination, and occlusion (Fig 2).

3 The Vehicle Re-Identification in Context Benchmark

3.1 Dataset Construction

We want to establish a realistic vehicle re-id evaluation benchmark with natural visual appearance characteristics and matching challenges (Sec 1). To this end, it is necessary to collect a large number of vehicle images/videos from wide-view traffic scenes. In the following, we describe the process of constructing the Vehicle Re-Identification in Context (VRIC) benchmark.

Source Video Data Given highly restricted access permission of typical surveillance video data, we propose to reuse existing vehicle related datasets publicly available in the research community.

In particular, we selected the UA-DETRAC object detection and tracking benchmark [24] as the source data of our VRIC benchmark, based on following considerations:

1. All videos were captured from the real-world traffic scenes (e.g. roads), reflecting realistic context for vehicle re-id.
2. It covers 24 different surveillance locations with diverse environmental conditions therefore offering a rich spectrum of test scenarios without bias towards particular viewing conditions.
3. It contains rich object and attribute annotations that can facilitate vehicle re-id labelling.

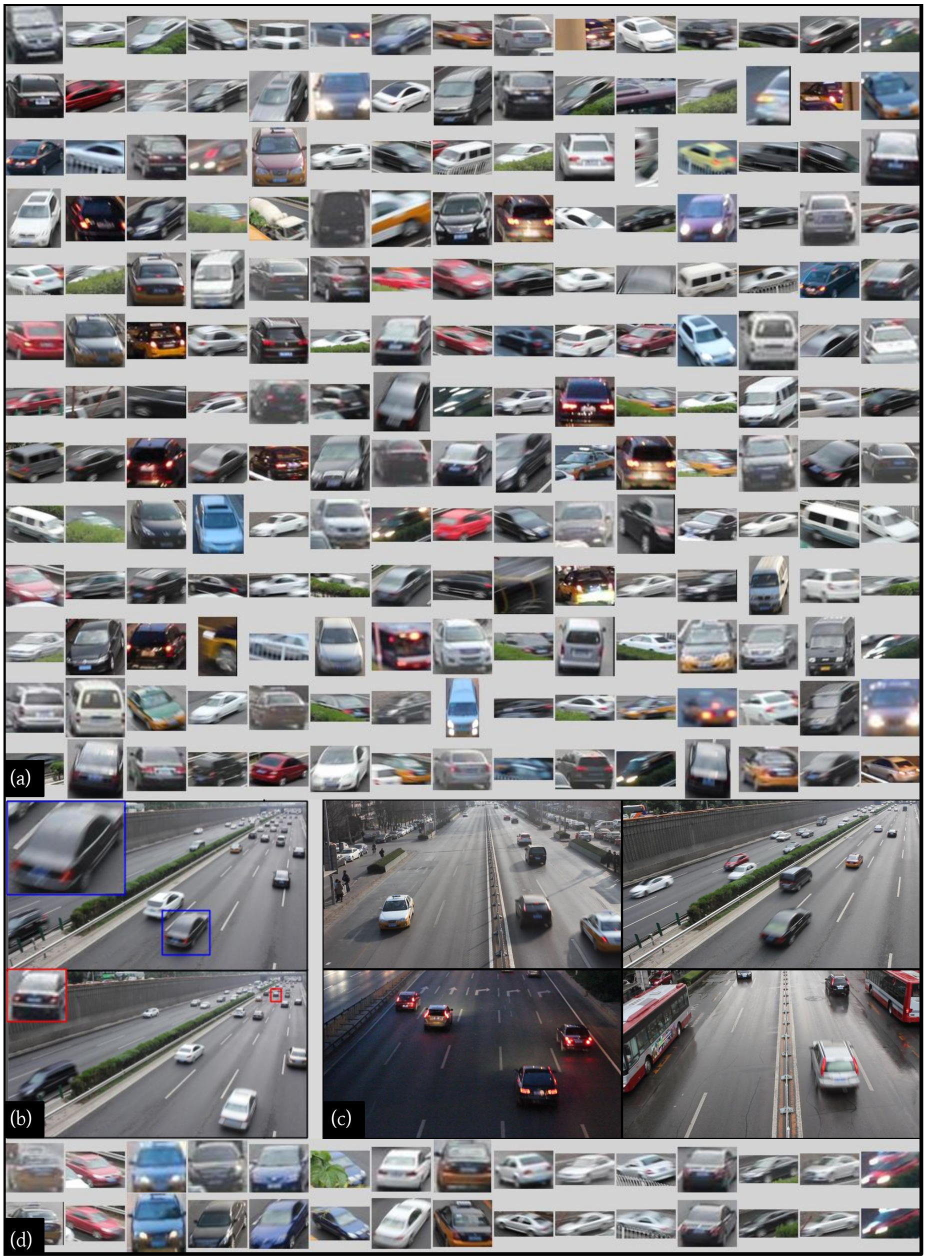

Fig. 2: Example vehicle bounding-box and whole scene images of the VRIC benchmark. **(a)** Samples of vehicle bounding-box images. **(b)** The *near* and *far* views in a wide-view traffic scene. **(c)** UA-DETRAC video scenes with different illumination due to changing weather conditions (sunny, cloudy and rainy) and time (day and night). **(d)** Vehicle matching pairs (each column) from some example test vehicle instances.

The UA-DETRAC videos were recorded at 25 frames per second (fps) with a frame resolution of 960×540 pixels (Fig 3). Samples of the whole scene images are shown in Fig 2(b,c).

Vehicle Image Filtering and Annotation. To construct a vehicle re-id dataset, we used 60 UA-DETRAC training videos with object bounding box annotations. For vehicle identity (ID) annotation, we started with assigning a unique label to each vehicle trajectory per UA-DETRAC video and then manually verified the ID duplication cases. Since all these raw videos were collected from different scenes and time durations, we found little duplicated trajectories in terms of identity. To ensure sufficient vehicle appearance variation, we throw away short trajectories with less than 20 frames and bounding boxes smaller than 24×24. By doing so, we obtained 5,622 vehicle IDs across all 60 videos.

In terms of vehicle instance resolution, the average image resolution of all 60,430 vehicle bounding-boxes is 69.8×107.5 pixels in width×height, with a variance of 32 to 280 pixels due to the unconstrained distances between vehicles and cameras. This presents inherently a multi-scale re-id matching challenge.

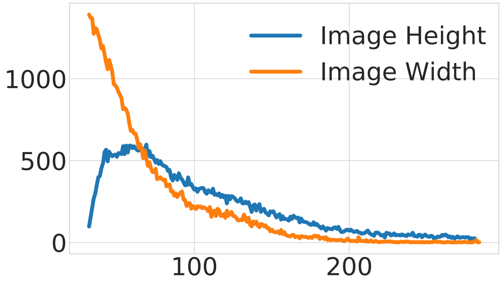

Fig. 3: Vehicle instance scale distributions in VRIC.

3.2 Evaluation Protocol

Data Split. For model training and testing using the VRIC dataset as a benchmark, we randomly split all 5,622 vehicle IDs into two non-overlapping halves: 2,811 for training, and 2,811 for testing. To remove data redundancy, we performed random frame-wise sub-sampling of the training trajectories. Since there is no cross-camera pairwise ID matches (UA-DETRAC is about single-camera object detection/tracking), we simulated cross-view variation by distant sampling between probe and gallery images.

In particular, we defined two pseudo views, *near* or *far*, for each video/camera and then built the probe/gallery sets from the test trajectories by randomly sampling each in two pseudo views. It is shown in Fig 2(b) that the *near* and *far* views present very different viewing conditions and hence allowing for a good simulation of two non-overlapping camera views. In this sense, VRIC contains a total 120 pseudo camera views from the 60 original camera views with unconstrained condition diversity.

We adopted the standard single-shot evaluation setting, i.e. one image per vehicle per view. From the above, we obtained 54,808/5,622 training/testing images for the VRIC benchmark. The data partition and statistics are summarised in Table 2.

Table 2: Data statistics and partition in VRIC.

Partition	All	Training Set	Test Set	
			Probe	Gallery
IDs	5,622	2,811	2,811	2,811
Images	60,430	54,808	2,811	2,811

Performance Metrics. For re-id performance measure, we used the *Cumulative Matching Characteristic* (CMC) rates [8]. The CMC is computed for each individual rank k as the cumulative percentage of the truth matches for probes returned at ranks $\leq k$. In practice, the Rank-1 rate is often used as a strong indicator of an algorithm's efficacy.

4 Deep Learning Multi-Scale Vehicle Representation

We aim to learn a deep representation model from a set of n vehicle images $\mathcal{I} = \{\boldsymbol{I}_i\}_{i=1}^{n}$ with the corresponding vehicle ID labels as $\mathcal{Y} = \{y_i\}_{i=1}^{n}$. These training images capture the visual appearance variations of n_{id} different IDs under multiple camera views, with $y_i \in [1, \cdots, n_{\mathrm{id}}]$. In typical surveillance scenes, vehicles are often captured at varying scales (resolutions), which causes significant inter-view feature representation discrepancy in re-id matching. In this work, we investigate this problem in vehicle re-id by exploring image pyramid representation [1,9].

Specifically, we exploit the potential of learning ID discriminative pyramidal representations originally designed for person re-id [3]. Our objective is to extract and represent complementary appearance information of vehicle ID from multiple resolution scales concurrently in order to optimise re-id matching under significant view changes. We call this model **Multi-Scale Vehicle Representation** (MSVR). Our approach differs notably from existing vehicle re-id models typically assuming single-scale representation learning.

MSVR Overview. The overall MSVR network design is depicted in Fig 4. Specifically, MSVR consists of $(m + 1)$ sub-networks: (1) m branches of sub-networks each for learning discriminative scale-specific visual features. Each

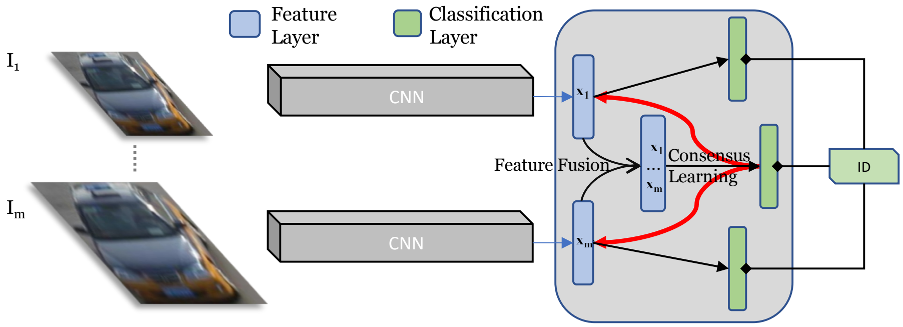

Fig. 4: Overview of Multi-Scale Vehicle Representation (MSVR) learning for discriminative vehicle re-id at varying spatial resolutions. MSVR learns vehicle re-id sensitive feature representations from image pyramid by an network architecture of multiple branches all of which are optimised concurrently (consensus feedback shown in red, see Eq. (4)) subject to the same ID label constraints. Importantly, an inter-scale interaction mechanism is enforced to further enhance the scale-generic feature learning.

branch has an identical structure. (2) One fusion branch for learning the discriminative integration of m scale-specific representations of the same vehicle image. To maximise the complementary advantage between different scales of feature representation in learning, we concurrently optimise per-scale discriminative representations with scale-specific and scale-generic (combined) learning subject to the same ID label supervision. Critically, we further propagate multi-scale consensus as feedback to regulate the learning of per-scale branches. Next, we detail three MSVR components: (1) Single-Scale Representation; (2) Multi-Scale Consensus; (3) Feature Regularisation.

(1) Single-Scale Representation. We exploit the MobileNet [7] to design single-scale branches due to its favourable trade-off between model complexity and learning capability. To train a single-scale branch, we use the softmax cross-entropy loss function to optimise vehicle re-id sensitive information from ID labels. Formally, we first compute the class posterior probability $\tilde{y}$ of a training image $\boldsymbol{I}$:

$$p(\tilde{y} = y|\boldsymbol{I}) = \frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{x})}{\sum_{k=1}^{n_\text{id}} \exp(\boldsymbol{w}_k^\top \boldsymbol{x})} \qquad (1)$$

where $\boldsymbol{x}$ and y refer to the feature vector and ground-truth label of $\boldsymbol{I}$, n_id the number of training IDs, and $\boldsymbol{w}_k$ the classifier parameters of class k. The training loss is then defined as:

$$\mathcal{L}_\text{ce} = -\log\Big(p(\tilde{y} = y|\boldsymbol{I})\Big) \qquad (2)$$

(2) Multi-Scale Consensus. We learn multi-scale consensus on vehicle ID classes between m scale-specific branches. We achieve this using joint-feature based classification. First, we obtain joint feature of different scales by vector fusion. In MobileNets, feature vectors are computed by global average pooling of the last CNN feature maps with dimension of 1024. Hence, this fusion produces a 1024×m-D feature vectors. We then use this combined features to perform classification for providing multi-scale consensus on the ID labels. We again adopt the cross-entropy loss (Eq (2)) as in single-scale representation learning.

(3) Feature Regularisation. We regularise the single-scale branches by multi-scale consensus for imposing interaction between different scale representations in model learning. Specifically, we propagate the consensus as an auxiliary *feedback* to regularise the learning of each single-scale branch concurrently. We first compute for each training sample a soft probability prediction (i.e. a consensus representation) $\tilde{P} = [\tilde{p}_1, \cdots, \tilde{p}_i, \cdots, \tilde{p}_{n_{\text{id}}}]$ as:

$$\tilde{p}_i = \tilde{p}(\tilde{y} = i | \boldsymbol{I}) = \frac{\exp(\frac{z_i}{T})}{\sum_k \exp(\frac{z_k}{T})}, \quad i \in [1, \cdots, n_{\text{id}}] \tag{3}$$

where z is the logit and T the temperature parameter (higher values leading to softer probability distribution). We empirically set $T = 1$ in our experiments. Then, we use the consensus probability $\tilde{P}$ as the *teacher* signal to guide the learning process of each single-scale branch (*student*). To quantify the alignment between these predictions, we use the cross-entropy measurement which is defined as:

$$\mathcal{H}(\tilde{P}, P) = -\frac{1}{n_{\text{id}}} \sum_{i=1}^{n_{\text{id}}} \big(\tilde{p}_i \ln(p_i) + (1 - \tilde{p}_i) \ln(1 - p_i)\big) \tag{4}$$

The objective loss function for each single-scale branch is then:

$$\mathcal{L}_{\text{scale}} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{H}(\tilde{P}, P) \tag{5}$$

where the hyper-parameter λ ($\lambda = 1$ in our experiments) is the weighting between two loss terms. $P = [p_1, \cdots, p_{n_{\text{id}}}]$ defines the probability prediction over all n_{id} identity classes by the corresponding single-scale branch (Eq. (1)). As such, each single-scale branch learns to correctly predict the true ID label of training sample ($\mathcal{L}_{\text{ce}}$) by the corresponding scale-specific representation and to match the consensus probability estimated based on the scale-generic representation ($\mathcal{H}$).

MSVR Deployment. In model test, we deploy the fusion branch's representation for multi-scale aware vehicle re-id matching. We use only a generic distance metric without camera-pair specific distance metric learning, e.g. the L2 distance. Based on the pairwise distance, we then return a ranking of gallery images as the re-id results. For successful tasks, the true matches for a given probe image are should be placed among top ranks.

5 Experiments

Datasets. For evaluation, in addition to the newly introduced VRIC dataset, we also utilised two most popular vehicle re-id benchmarks. The **VehicleID** [14] dataset provides a training set with 113,346 from 13,164 IDs and a test set with 19,777 images from 2,400 identities. It adopts the single-shot re-id setting, with only one true matching for each probe. Following the standard setting, we repeated 10 times of randomly selected probe and gallery sets in our experiments. The **VeRi-776** dataset [16] has 37,778 images of 576 IDs in training set and 200 IDs in test set. The standard probe and gallery sets consist of 1,678 and 11,579 images, respectively. The data split statistics are summarised in Table 3.

Table 3: Data split of vehicle re-id datasets evaluated in our experiments.

Dataset	Training IDs / Images	Probe IDs / Images	Gallery IDs / Images
VehicleID[14]	13,164 / 113,346	2,400 / 17,377	2,400 / 2,400
VeRi-776[16]	576 / 37,778	200 / 1,678	200 / 11,579
VRIC (**Ours**)	2,811 / 54,808	2,811 / 2,811	2,811 / 2,811

Performance Metrics. For VehicleID and VRIC, we used the CMC measurement to evaluate re-id performance. For VeRi-776, we additionally adopted the *mean Average Precision* (mAP) due to its multi-shot nature in the gallery of the test data. Specifically, for each probe, we compute the area under its Precision-Recall curve, i.e. Average Precision (AP). The mAP is then computed as the mean value of APs for all probes. This metric considers both precision and recall performance, and hence providing a more comprehensive evaluation.

Implementation Details. In the MSVR model, we used 2 resolution scales, 224×224 and 160×160. We adopted the ADAM optimizer and set the initial learning rate to 0.0002, the weight decay to 0.0002, the β_1 to 0.5, the mini-batch size to 8, the max-iteration to 100,000. Model initialization was done with ImageNet [4] pretrained weights. The data augmentation includes random cropping and horizontal flipping.

Evaluation. Table 4 compares MSVR with state-of-the-art methods on three benchmarks. We make these main observations as follows:
(1) Under the standard visual appearance based evaluation setting (the top part), MSVR outperforms all other competitors with large margins − MSVR surpasses the best competitor in Rank-1 rate by 24.38 % (88.56-64.18) on VeRi-776, 24.82% (62.02-38.20) on VehicleID, and 16.73% (46.61-30.55) on VRIC. This demonstrates the consistent superiority of MSVR over alternative methods in vehicle re-id, showing the importance in modelling multi-scale representation

for vehicle re-id.

(2) Benefited from more training data plus space-time contextual knowledge and fine-grained local key-point supervision, the OIFE model achieves the best performance on VeRi-776. However, such advantages from additional data and knowledge representation is generically beneficial to all models including the MSVR when applied.

(3) We carefully reproduced two very recent methods, OIFE(Single-Branch) [23] and Siamese-Visual [19], and obtained inconsistent results compared to the reported performances of these two models. In particular, the performance of OIFE(Single-Branch) decreases on VeRi-776 and VehicleID. This is mainly due to that the original results are based on a larger multi-source training set with 225,268 training images of 36,108 IDs (from VehicleID [14], VeRi-776 [16], Box-Cars [20] and CompCars [27]), *versus* the standard 100,182 training images of 13,164 IDs on VehicleID, *i.e.* 2.2 times more training images and 2.7 times more training ID labels, and the standard 37,778 training images of 576 IDs on VeRi-776, *i.e.* 6.0 times more training images and 62.7 times more training ID labels, respectively. In contrast, the result of Siamese-Visual (ResNet50 based) increases on VeRi-776. It is worth pointing out that we trained this model using the cross-entropy classification loss and cannot make it converge with pairwise inner-product loss.

Table 4: Comparative vehicle re-id results on three benchmarking datasets. Upper part of table lists methods trained with only the images available from the respective datasets for fair comparison of the methods; lower part lists methods trained with additional datasets and/or labels. *: By our reimplementation. **E**: Extra information and annotation, *e.g.* number plates, local key-points, space-time prior knowledge. **M**: Multiple vehicle re-id and classification datasets are combined for training. †: Result from [23].

Method	Notes	VeRi-776 [16]		VehicleID [14]		VRIC		Publication
		Rank-1	mAP	Rank-1	Rank-5	Rank-1	Rank-5	
LOMO [13]		25.33	9.64	-	-	-	-	CVPR'15
FACT [15]		50.95	18.49	-	-	-	-	ICME'16
Mixed Diff + CCL [14]		-	-	38.20	50.30	-	-	CVPR'16
Siamese-Visual [19]		41.12	29.40	-	-	-	-	ICCV'17
Siamese-Visual [19]	*	64.18	31.54	36.83	57.97	30.55	57.30	ICCV'17
OIFE(Single Branch) [23]	*	60.13	31.81	32.86	52.75	24.62	50.98	ICCV'17
MSVF		**88.56**	**49.30**	**63.02**	**73.05**	**46.61**	**65.58**	**Ours**
KEPLER [17] †	M	68.70	33.53	45.40	68.90	-	-	TIP'15
FACT + Plate + Space-Time [16]	E	61.44	27.77	-	-	-	-	ECCV'16
Siamese-CNN + Path-LSTM [19]	E	83.49	**58.27**	-	-	-	-	ICCV'17
OIFE(Single Branch) [23]	M	88.66	45.50	63.20	80.60	-	-	ICCV'17
OIFE(4Views) [23]	ME	89.43	48.00	**67.00**	**82.90**	-	-	ICCV'17
OIFE(4Views + Space-Time) [23]	ME	**92.35**	51.42	-	-	-	-	ICCV'17

Further Analysis. Table 5 compares the performances of a single-scale and a multi-scale feature representations of the MSVR model. It is evident that the multi-scale representation learning with MSVR has performance benefit across all three datasets with varying resolution scale changes. This shows that the overall effectiveness of MSVR in boosting vehicle re-id matching performance. Moreover, the model performance gain on VRIC is the largest, which is consistent with the more significant scale variations exhibited in the VRIC vehicle images (Fig 1 and Table 1).

Table 5: Comparing single-scale and multi-scale representations of MSVR. Gain is measured as the performance difference of MSVR over the *mean* of single-scale variants.

Dataset	VeRi-776 [16]		VehicleID [14]		VRIC	
Metrics (%)	Rank-1	mAP	Rank-1	Rank-5	Rank-1	Rank-5
Scale-224	88.37	47.37	62.80	72.54	43.55	61.88
Scale-160	87.43	46.81	60.29	71.15	43.62	62.77
MSVR	**88.56**	**49.30**	**63.02**	**73.05**	**46.61**	**65.58**
Gain (%)	+0.76	+2.11	+1.47	+1.20	**+3.02**	**+3.25**

6 Conclusion

In this work we introduced a more realistic and challenging vehicle re-identification benchmark, Vehicle Re-Identification in Context (VRIC), to enable the design and evaluation of vehicle re-id methods to more closely reflect real-world application conditions. VRIC is uniquely characterised by unconstrained vehicle images from large scale, wide scale traffic scene videos inherently exhibiting variations in resolution, illumination, motion blur, and occlusion. This dataset provides a more realistic and truthful test and evaluation of algorithms for vehicle re-id "in-the-wild". We further investigated a multi-scale learning representation by exploiting a pyramid based deep learning method. Experimental evaluations demonstrate the effectiveness and performance advantages of our multi-scale learning method over the state-of-the-art vehicle re-id methods on three benchmarks VeRi-776, VehicleID, and VRIC.

Acknowledgements

This work was partially supported by the Royal Society Newton Advanced Fellowship Programme (NA150459), Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149), Vision Semantics Ltd, and SeeQuestor Ltd.

References

1. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA Engineer **29**(6), 33–41 (1984)
2. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
3. Chen, Y., Zhu, X., Gong, S., et al.: Person re-identification by deep learning multi-scale representations. In: Workshop of IEEE International Conference on Computer Vision (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
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 (2012)
6. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person re-identification. Springer (January 2014)
7. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
8. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2006)
10. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
11. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: International Joint Conference of Artificial Intelligence (2017)
12. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
13. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
14. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
15. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: IEEE International Conference on Multimedia and Expo (2016)
16. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European Conference on Computer Vision (2016)
17. Martinel, N., Micheloni, C., Foresti, G.L.: Kernelized saliency-based person re-identification through multiple metric learning. IEEE Transactions on Image Processing **24**(12), 5645–5658 (2015)

18. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499 (2016)
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: IEEE International Conference on Computer Vision (2017)
20. Sochor, J., Herout, A., Havel, J.: Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3006–3015 (2016)
21. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: IEEE International Conference on Computer Vision (2017)
22. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701–1708 (2014)
23. Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., Yan, J., Wang, S., Li, H., Wang, X.: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: IEEE International Conference on Computer Vision (2017)
24. Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M., Qi, H., Lim, J., Yang, M., Lyu, S.: UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. arXiv **abs/1511.04136** (2015)
25. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2016)
26. Yan, K., Tian, Y., Wang, Y., Zeng, W., Huang, T.: Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In: IEEE International Conference on Computer Vision. pp. 562–570 (2017)
27. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
28. Zapletal, D., Herout, A.: Vehicle re-identification for automatic video traffic surveillance. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1568–1574 (2016)
29. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: European Conference on Computer Vision. pp. 94–108 (2014)
30. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)