# Large-scale vehicle re-identification in urban surveillance videos

**Conference Paper** · July 2016

DOI: 10.1109/ICME.2016.7553002



**4 authors**, including:

Xinchen Liu
JD AI Research
**16** PUBLICATIONS **438** CITATIONS
SEE PROFILE

Wu Liu
AI Research of JD.com
**62** PUBLICATIONS **1,038** CITATIONS
SEE PROFILE

# LARGE-SCALE VEHICLE RE-IDENTIFICATION IN URBAN SURVEILLANCE VIDEOS

*Xinchen Liu, Wu Liu, Huadong Ma, Huiyuan Fu*

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia
Beijing University of Posts and Telecommunications, Beijing 100876, China
xinchenliu@bupt.edu.cn, liuwu@live.cn, {mhd, fhy}@bupt.edu.cn

## ABSTRACT

Vehicle, as a significant object class in urban surveillance, attracts massive focuses in computer vision field, such as detection, tracking, and classification. Among them, vehicle re-identification (Re-Id) is an important yet frontier topic, which not only faces the challenges of enormous intra-class and subtle inter-class differences of vehicles in multi-cameras, but also suffers from the complicated environments in urban surveillance scenarios. Besides, the existing vehicle related datasets all neglect the requirements of vehicle Re-Id: 1) massive vehicles captured in real-world traffic environment; and 2) applicable recurrence rate to give cross-camera vehicle search for vehicle Re-Id. To facilitate vehicle Re-Id research, we propose a large-scale benchmark dataset for vehicle Re-Id in the real-world urban surveillance scenario, named "VeRi". It contains over 40,000 bounding boxes of 619 vehicles captured by 20 cameras in unconstrained traffic scene. Moreover, each vehicle is captured by 2∼18 cameras in different viewpoints, illuminations, and resolutions to provide high recurrence rate for vehicle Re-Id. Finally, we evaluate six competitive vehicle Re-Id methods on VeRi and propose a baseline which combines the color, texture, and high-level semantic information extracted by deep neural network.

***Index Terms***— Vehicle Re-identification, Benchmark, Recurrence Rate, Deep Neural Network

## 1. INTRODUCTION

This paper focuses on **vehicles re-identification** (Re-Id) in urban surveillance scenario, which has been largely neglected by vision community compared to other objects like persons [1, 2]. Vehicle, such as cars, trucks, and buses, is a rich object class which can potentially bring many interesting research topics [3]. Previous vehicle related research has focused on vehicle detection, tracking and classification [4, 5].

This work is supported by the National Natural Science Foundation of China under Grant No. 61332005 and No. 61402048; The Funds for Creative Research Groups of China under Grant No. 61421061; The Cosponsored Project of Beijing Committee of Education; The Beijing Training Project for the Leading Talents in S&T (ljrc 201502); The NSFC-Guangdong Joint Fund (U1501254).

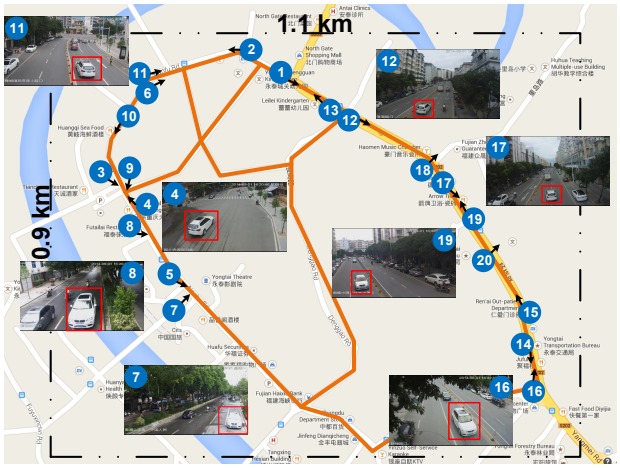



**Fig. 1**. The urbran surveillance environments and cameras distribution for the VeRi dataset. The blue circles and arrows denote the locations and directions of cameras. The surveillance camera views illustrate an example of vehicle Re-Id: the same white BMW SUV captured by the surveillance network.

While vehicle Re-Id, comparing to person Re-Id [6, 7], is still a frontier area. As shown in Fig. 1, the task of vehicle Re-Id is, given a probe vehicle image, to search in a database for images that contain the same vehicle captured by multiple cameras. Vehicle Re-Id can play great roles in intelligent transportation [8], urban computing [9], and intelligent surveillance [10], which can quickly discover, locate, and track the target vehicles.

Different from vehicle detection, tracking or classification, vehicle Re-Id in urban surveillance video can be found as a near duplicate image retrieval (NDIR) problem [11]. For the probe vehicle image, the vehicles with the similar appearance need to be found firstly. However, more than conventional NDIR, only depending on appearance can hardly give optimal vehicle Re-Id results as the large intra-class difference of the same vehicle in different cameras, and subtle inter-class differences between different vehicles in the same view (some examples can be found in Fig.2(b)). Unquestionably, license plate is a significant cue for vehicle Re-Id. Nonetheless, it may not work well in unconstrained surveillance scenes due to the various illuminations, viewpoints, and occlusions.

In particular, the lack of proper datasets is another key problem to hinder the development of vehicle Re-Id. Most existing vehicle related datasets are usually applied for classification. For example, the wheeled vehicle synset of the ImageNet [12] contains 1,537 vehicle images. Recently, Yang *et al.* [3] have proposed a cars dataset, "CompCars", which contains web-nature images crawled from the web and surveillance images captured in the front view. It is mainly used for fine-grained categorization and attribute prediction, which also neglect the requirements of vehicle Re-Id. A comprehensive vehicle Re-Id dataset must not only contain massive vehicles captured by cameras in real-world traffic environment, but also include enough recurrence rate. The recurrence rate means that the same vehicle must appear in different cameras with different viewpoints and backgrounds, which guarantees to give cross-camera vehicle search for vehicle Re-Id.

To facilitate future research of vehicle Re-Id, in this paper, we build a large-scale benchmark dataset for vehicle Re-Id (VeRi) in the real-world urban surveillance scenario, and propose a benchmark vehicle Re-Id method by evaluating six different vehicle Re-Id schemes on the dataset. The featured properties of VeRi include:

1. It contains over 40,000 images of 619 vehicles captured by 20 cameras covering an $1.0km^2$ area in 24 hours (as shown in Fig. 1), which makes the dataset scalable enough for vehicle Re-Id and other related research.
2. The images are captured in a real-world unconstrained surveillance scene and labeled with varied attributes, e.g. BBoxes, types, colors, and brands. So complicated models can be learnt and evaluated for vehicle Re-Id.
3. Each vehicle is captured by $2 \sim 18$ cameras in different viewpoints, illuminations, resolutions, and occlusions, which provides high recurrence rate for vehicle Re-Id in practical surveillance environment.

Some examples in VeRi can be found in Fig. 2. The dataset contains vehicles with various types and models (shown in Fig. 2(a)), which brings great challenges for the vehicle Re-Id task. More importantly, as shown in Fig. 2(b), we reconstruct the real-world challenges of enormous intra-class difference and subtle inter-class difference for vehicle Re-Id.

The other important contribution in this paper is that we give a benchmark vehicle Re-Id method on the proposed dataset, which fuses the vehicle's color, texture, and high-level semantic features for vehicle Re-Id. First of all, the color features are extracted by the benchmark method in person Re-Id [7], which applies Bag-of-Words model with Color Name features. For texture feature, we used the classical BoW with Scale-Invariant Feature Transform (SIFT) descriptor in near duplicate image retrieval [13]. Recently, deep Convolutional Neural Network (CNN) has achieved excellent performance in many computer vision tasks [14–17]. The vehicle's semantic features are extracted by a fine-tuned deep CNN, which are proposed as a benchmark method in [3] for car model classification and verification. On the VeRi, we comprehensively evaluate the effects of color, texture, and high-level semantic features, and demonstrate that the combination of the three features can give the best performance.

## 2. RELATED WORKS

As vehicle has many special characters such as the diversity of brands, designs, and functions, many researches have focused on vehicle from different research interests. One hot topic is the fine-grained categorization. For example, Yang *et al.* [3] used CNN to extract visual features from images of entire and part vehicles for fine-grained classification and attribute prediction. 3D model based approaches was used to match and fit the 2D image for car model recognition [18]. However, vehicle Re-Id is still on its early stage, with a handful of related works. Among them, Feris *et al.* [4] proposed a vehicle detection and attribute-based retrieval system, in which vehicles are searched by attributes like colors and types coarsely.

As a similar task for different objects, person Re-Id has been a popular topic. For example, with color information, Zheng *et al.* [7] adopted an unsupervised Bag-of-Words model with color features, which yields competitive accuracy for person Re-Id. With texture features, Farenzena *et al.* [1] proposed the Symmetry-Driven Accumulation of Local Features for person Re-Id. Recently, Li *et al.* [6] proposed a deep filter pairing neural network to jointly optimize all pipeline of person re-id and achieved competitive accuracy. These works proved that both the global features such as color and local texture descriptors are effective to represent objects. The recent deep neural network can catch consistent details of objects and provide an end-to-end framework for person re-id. These motivations can be exploited for vehicle re-id.

In recent years, many vehicle related datasets have been released for classification. For example, Yang *et al.* [3] proposed a comprehensive cars dataset, named "CompCars", including images from two sources. The first is crawled from the web containing 1,687 car models covering most of the commercial car models in the recent years. Nonetheless, the second parts of images are captured by the surveillance camera only in the front view. Moreover, there is no cross-camera correlations for the vehicles in this kind of datasets, which cannot give enough recurrence rate for vehicle Re-Id. Therefore, this dataset is not suitable for vehicle Re-Id. Closely related to vehicle Re-Id, many person Re-Id datasets have been constructed [6,7], which can give important references for vehicle Re-Id dataset building. As an example, Zheng *et al.* [7] constructed a large-scale person Re-Id dataset, Market-1501, containing 1501 identities captured by 6 cameras. It is now the largest person Re-Id dataset with 32,668 images, 500K distractors, and 3,368 queries.

Inspired by the above works, we construct the VeRi, which is the most comprehensive dataset for vehicle Re-Id with many useful features, e.g. large-scale data volume, diverse labels, cross-camera recurrence, and real-world environment settings. Furthermore, a state-of-the-art is proposed on this dataset with fusing of color, texture and high-level semantics based features for vehicle Re-Id.

(a) (b)

**Fig. 2**. Vehicle samples selected from the VeRi dataset. All images are normalized to the same size. (**a**) Diversity of vehicle colors and types. (**b**) Variation of the viewpoints, illuminations, resolutions, and occlusions for the vehicles in different cameras.

## 3. DATESET CONSTRUCTION

In this section, the dataset construction is introduced in detail: the environments of the surveillance scene and device settings, the annotation procedures, and the schemes used in the construction of the VeRi dataset.

### 3.1. Environments and Device Settings

To collect high-quality videos in real-world surveillance scene, we select 20 cameras deployed along a circular road of an $1.0km^2$ area as shown in Fig. 1. The scenes of the cameras include two-lane roads, four-lane roads, and crossroads. All cameras are set to $1920 \times 1080$ resolution and 25 fps. The cameras are deployed with arbitrary orientations and tilt angles. Besides, there are overlaps for part of the cameras.

### 3.2. Construction Procedure

With the 20 cameras, we record continuous 24-hour videos and obtain 1.43 TB data. Then we select the videos from 4:00 p.m. to 5:00 p.m., which contains 34.2 GB videos after simple transcoding and compression. Considering efficiency and quality, we sample one in every five frames from the videos to obtain near 360,000 frames in all. The dataset construction procedure is divided into three main steps.

#### *3.2.1. Bounding Box and Track Annotation*

We first annotate bounding boxes of vehicles in each frame. A bounding box (BBox) is a rectangle that surrounds a whole vehicle body. Only the moving vehicles are labeled with BBoxes. The BBoxes smaller than $64 \times 64$ will be dropped. The second task is to create the track, which is the trajectory of a vehicle captured by a camera at a time. The BBoxes belonging to a track are clustered together to guarantee each query has multi-ground truth during search. The camera ID and timestamp (frame ID) are reserved with tracks for further annotation. At last, for each track, at most six BBox images are saved considering the resolution of the images and the diversity of the vehicle appearances. After this step, we obtain about 9,000 tracks and 50,000 BBoxes.

#### *3.2.2. Color and Type Categorization*

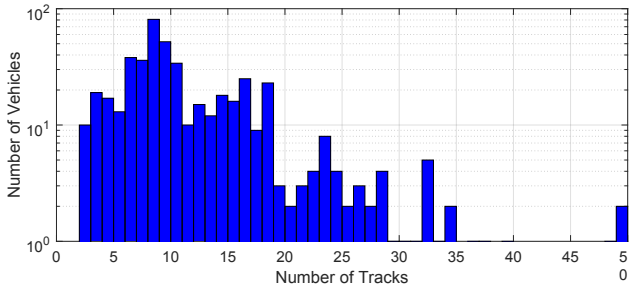
As a preprocessing step for cross-camera correlation, we categorize all tracks under each camera into ten colors and nine types. According to [3], we define a list of colors including black, gray, white, red, green, orange, yellow, golden, brown, and blue. Then for each color, each track is classified into one of ten vehicle types, i.e. sedan, SUV, hatchback, MPV, van, pickup, bus, truck, and estate car. The color and type labels are assigned by three workers using a majority vote.

#### *3.2.3. Cross-Camera Vehicle Correlation*

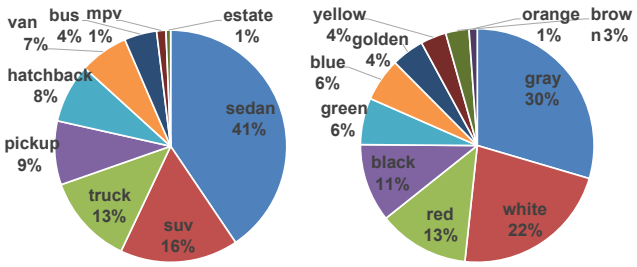
The cross-camera vehicle correlation is the most important annotation for vehicle Re-Id. In other words, we need to label the ground truth for vehicle Re-Id. Meanwhile it is the most difficult and time-cost task as we should correlate the tracks of the same vehicle in thousands tracks under all cameras. In the labeling, some strategies are used to reduce the workload. With the hierarchical categorization of tracks, we only consider the tracks with similar colors and types in each camera. Then the appearance and license plate (if it can be recognized) are used to identify the same vehicle. Besides, the temporal information of the tracks and the spatial correlation of cameras can also assist the task. In spite of this, the cross-camera vehicle correlation task still costs more than one-month workloads of six fulltime workers. Finally, we obtain 619 vehicles after filtering the vehicles that appear only once.

### 3.3. VeRi Dataset at a Glance

This section looks into the typical properties in VeRi dataset. The VeRi dataset contains 619 vehicles, 7,021 tracks and




(a) Distribution of vehicle track numbers.




(b) Distribution of vehicle colors and types.

**Fig. 3**. The key characters of the proposed VeRi dataset.

40,395 BBox images. This make our dataset scalable enough for vehicle Re-Id task. Furthermore, each vehicle is captured by at least two cameras. The distributions of vehicle track numbers is shown in Fig. 3(a). It shows that most vehicles have more than five tracks over the whole camera network, which guarantees the recurrence rate of vehicles.

Because of the natural environments and unconstrained traffic settings, our dataset can be used for many real-world surveillance applications. This also make VeRi a very challenging dataset for vehicle Re-Id. One of the challenges is the variable lightning conditions as shown in Fig. 2(b). The color of the vehicle would change dramatically due to the changing illumination, shadows and the specular reflection of car body. Another challenge is to identify the same vehicle from different views. As shown in Fig. 1(b), it it very difficult to find the same car with different viewpoints. While from the same view, two vehicles may look very similar in appearance especially they belong to the same model. Furthermore, the different distances of the cameras also make the resolutions different for the same vehicle.

As the target of vehicle Re-Id is to retrieve all the appeared images for the query vehicle, the cross-camera correlations for the same vehicle are the key property in the VeRi dataset. All tracks and BBoxes with the same vehicle are clustered together and labeled with a vehicle ID. When one query comes, the images with the same vehicle ID but different camera ID are its ground truth. Moreover, our dataset is labeled with varied colors and types. The distributions of vehicle colors and types are shown in Fig. 3(b).

## 4. THE VEHICLE RE-ID METHODS

In this section, we implement six competitive vehicle Re-Id methods as comparative methods, and propose a baseline which combines the color, texture, and high-level semantic information learned by deep neural network. The details of the color, texture, and semantic features can by found as follows.

**1. Texture based feature (BOW-SIFT).** We first resize the vehicle image to $64 \times 128$. Then the SIFT descriptors [19] are extracted as the local texture features of vehicles. Next, BOW model is used to quantize the features due to its accuracy and efficiency in NDIR. The model is implemented as in [13] with the codebook trained on our training set by hierarchical k-means, $k = 10,000$. This model is used as the texture based feature for vehicle Re-Id.

**2. Color based feature (BOW-CN).** This is a benchmark method in person Re-Id, which applies BOW with Color Name (CN) feature [20]. The CN feature is adopted as it is an effective and robust image feature for outdoor environments. The vehicle image is resized to $64 \times 128$, then the CN feature is densely sampled for each $4 \times 4$ patch with the 4-pixel sampling step. The codebook of BOW is trained on our training set using standard k-means, $k = 350$. Moreover, we also adopt the avgIDF, weak geometric constrains, and gaussian background suppression techniques in [7]. With the weak geometric constrains, each vehicle image is partitioned into 16 strips. For each strip, the BOW-CN feature is computed separately. Finally, by concatenating all features of 16 strips, we obtain a 5600-D BOW-CN feature as the color based feature.

**3. Semantic feature extracted by deep neural network.** We explore two semantic features learned by Convolutional Neural Networks. **1) AlexNet.** We adopt the AlexNet [14] used for ImageNet classification task. The semantic feature of AlexNet is learned from 1,000 classes of general objects. A 4096-D image feature is extracted from the FC6 layer of AlexNet. **2) GoogleLeNet.** This model is the GoogLeNet [21] fine-tuned as [3] on the CompCars dataset. In [3], the model is trained on the images of the entire and part cars to detect the detailed attributes of vehicles, such as the number of doors, the shape of lights, the number of seats, and type of cars. We can employ the feature extracted from the model as the semantic features, which catches the high-level semantic information of vehicles.

**4. Feature Fusion.** We also exploit the fusion of multiple features, i.e. color, texture, and semantic feature. **1) AlexNet + BOW-CN.** This model uses the early fusion scheme which directly concatenates the semantic features extracted by AlexNet with the BOW-CN feature. **2) Fusion of Attributes and Color feaTures (FACT).** The proposed FACT model uses the late fusion scheme. We first obtain the rank scores of all test images with the BOW-SIFT, BOW-CN, and the semantic feature learned by GoogLeNet separately. For all experiments, the rank scores are calculated by the Euclidean distance. Then the three types of scores are summed

**Table 1**. Comparison of the vehicle Re-Id methods on VeRi.

| methods | mAP | HIT@1 | HIT@5 |
|---|---|---|---|
| BOW-SIFT [13] | 1.85 | 2.81 | 5.82 |
| BOW-CN [7] | 13.95 | 46.56 | 61.88 |
| AlexNet [14] | 9.69 | 42.39 | 55.09 |
| AlexNet + BOW-CN | 13.92 | 42.68 | 58.87 |
| GoogLeNet [3] | 17.88 | 58.87 | 74.10 |
| FACT | **19.92** | **59.65** | **75.27** |

with different weights. According to the experiment results of the single-feature models, the fusion weights are set to 0.1, 0.2, and 0.7 respectively due to their individual performances.

## 5. EXPERIMENTS

### 5.1. Experiment settings

Before conducting experiments, we randomly divide the VeRi dataset into two parts for training and testing respectively. The training set has 499 vehicles with 33,195 images and the testing set contains 120 vehicles with 7,020 images. For each vehicle in the testing set, we select one image from each camera as the query and obtain 1,031 query images in all. In the evaluation, the cross-camera search is performed, which means we use one image of a vehicle from one camera to search for the images of the same vehicle from other cameras. To evaluate the performance of the proposed algorithms in Section 4, we use the Cumulative Matching Characteristic (CMC) curve which is widely used in person Re-Id [7]. It computes the probability that a query vehicle appears in different-sized candidate lists. Moreover, we also use mean average precision (mAP), HIT@1, and HIT@5 to evaluate the overall performance.

### 5.2. Results Analysis

Table 1 illustrates the mAP, HIT@1, and HIT@5 of different methods. The CMC curves are plotted in Fig. 4. BOW-SIFT which is used in NDIR has the worst performance. The reason is that different vehicles in the same model always have similar textures, but the same vehicle may have different textures as the different capture viewpoints. The BOW-CN model achieves competitive accuracy which also yields good performance in the person Re-Id [7]. So color features are effective descriptors for vehicle Re-Id. The AlexNet model for object classification is not learned specifically for vehicles and obtains relatively low accuracy. The fine-tuned GoogLeNet model achieves better results among single-feature models. It proves that the fine-tuned CNN can catches more detailed features as it is learned from images of entire and part cars to recognize the detail attributes of vehicle, such as the number of doors, the shape of lights, and type of cars. So it can distinguish vehicles by attributes based features, which can be seen

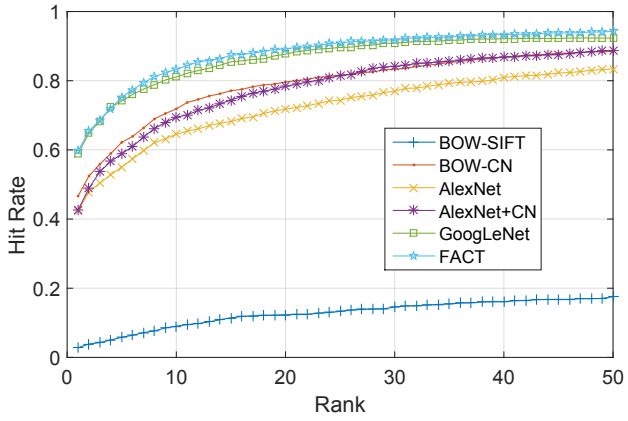



**Fig. 4**. Comparison of the vehicle Re-Id methods on VeRi in terms of CMC curves.

as the high-level semantic information. Moreover, the fusion of texture, color, and the semantics based features learned by CNN yields better performance than single-feature models, which demonstrates the effectiveness of the proposed FACT method.

Some sample results of the FACT on VeRi dataset can be found in Fig. 5. From the Fig 5 (a) and (b), we can find FACT can give optimal accuracy for vehicle Re-Id. However, there are also challenges in the VeRi dateset such as the similarity of the vehicle appearances as shown in Fig. 5(c). Besides, the uncertain illumination and the specular reflection make different vehicles have similar color as shown in Fig. 5(d).

### 5.3. Discussion and Perspectives

From the results on the constructed dataset, we can also find that: **1)** Except visual similarities search, we must consider the license plate for accurate vehicle Re-Id. Especially, visual similarities can only help us quickly find the similar objects. More specific, we need the specific ID to find the exactly matched vehicle. Therefore, the next key research topic for the proposed method in the future is using vehicle plate matching to improve the vehicle Re-Id performance. **2)** Except the visual similarities, for vehicle Re-Id, there are many additional and useful information, such as temporal information and road network information. How to fusing the additional information to improve the accuracy of vehicle Re-Id is also an important research topic in the future.

## 6. CONCLUSION AND FUTURE WORKS

This paper proposes a large-scale image dataset for vehicle Re-Id, named "VeRi". It contains over 40,000 annotated BBoxes of 619 vehicles captured by 20 cameras in unconstrained urban surveillance environments. Each vehicle is captured by 2$\sim$18 cameras in different viewpoints, illuminations, resolutions, and occlusions to provide high recurrence


**Fig. 5**. Sample results of the FACT on VeRi dataset. The query is in blue box and the top-16 results are listed. The true positive results are in green box, otherwise red.

rate for vehicle Re-Id. Finally, a vehicle Re-Id method which combines the texture, colors, and high-level attributes information is proposed as a baseline on the VeRi dataset. In the future, we will try to utilize FACT as a coarse search, then exploit to combine license plate recognition to further improve the performance of vehicle Re-Id.

## 7. REFERENCES


[1] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[2] Wang Junqiang and Huadong Ma, "Pedestrian detection with geometric context from a single image," in *ACM MM*, 2011, pp. 1225–1228.

[3] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "A large-scale car dataset for fine-grained categorization and verification," in *CVPR*, 2015, pp. 3973–3981.

[4] Rogerio Schmidt Feris, Behjat Siddiquie, James Petterson, Yun Zhai, Ankur Datta, Lisa M Brown, and Sharath Pankanti, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE TMM*, vol. 14, no. 1, pp. 28–42, 2012.

[5] Bogdan C Matei, Harpreet S Sawhney, and Supun Samarasekera, "Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features," in *CVPR*, 2011, pp. 3465–3472.

[6] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.

[7] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, Jiahao Bu, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[8] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen, "Data-driven intelligent transportation systems: A survey," *IEEE TITS*, vol. 12, no. 4, pp. 1624–1639, 2011.

[9] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang, "Urban computing: concepts, methodologies, and applications," *ACM TIST*, vol. 5, no. 3, pp. 38, 2014.

[10] Wu Liu, Yongdong Zhang, Sheng Tang, Jinhui Tang, Richang Hong, and Jintao Li, "Accurate estimation of human body orientation from rgb-d sensors," *IEEE TCYB*, vol. 43, no. 5, pp. 1442–1452, 2013.

[11] Tao Mei, Yong Rui, Shipeng Li, and Qi Tian, "Multimedia search reranking: A literature survey," *ACM CSUR*, vol. 46, no. 3, pp. 38, 2014.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[13] Liang Zheng, Shengjin Wang, Wengang Zhou, and Qi Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *CVPR*, 2014, pp. 1963–1970.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[15] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu, "Siamese neural network based gait recognition for human identification," in *ICASSP*, 2016.

[16] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *CVPR*, 2015, pp. 3707–3715.

[17] Xishan Zhang, Hanwang Zhang, Yong Dong Zhang, Yang Yang, Meng Wang, Huanbo Luan, Jin Tao Li, and Tat-Seng Chua, "Deep fusion of multiple semantic cues for complex event recognition," *IEEE TIP*, vol. 25, no. 3, pp. 1033–1046, 2016.

[18] Zhaoxiang Zhang, Tieniu Tan, Kaiqi Huang, and Yunhong Wang, "Three-dimensional deformable-model-based localization and recognition of road vehicles," *IEEE TIP*, vol. 21, no. 1, pp. 1–13, 2012.

[19] David. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[20] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus, "Learning color names for real-world applications," *IEEE TIP*, vol. 18, no. 7, pp. 1512–1523, 2009.

[21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.