PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance

Xinchen Liu, *Student Member, IEEE*, Wu Liu, *Member, IEEE*, Tao Mei, *Senior Member, IEEE*, and Huadong Ma, *Senior Member, IEEE*

***Abstract*—Compared with person reidentification, which has attracted concentrated attention, vehicle reidentification is an important yet frontier problem in video surveillance and has been neglected by the multimedia and vision communities. Since most existing approaches mainly consider the general vehicle appearance for reidentification while overlooking the distinct vehicle identifier, such as the license plate number, they attain suboptimal performance. In this paper, we propose PROVID, a PROgressive Vehicle re-IDentification framework based on deep neural networks. In particular, our framework not only utilizes the multimodality data in large-scale video surveillance, such as visual features, license plates, camera locations, and contextual information, but also considers vehicle reidentification in two progressive procedures: coarse-to-fine search in the feature domain, and near-to-distant search in the physical space. Furthermore, to evaluate our progressive search framework and facilitate related research, we construct the VeRi dataset, which is the most comprehensive dataset from real-world surveillance videos. It not only provides large numbers of vehicles with varied labels and sufficient cross-camera recurrences but also contains license plate numbers and contextual information. Extensive experiments on the VeRi dataset demonstrate both the accuracy and efficiency of our progressive vehicle reidentification framework.**

***Index Terms*—Progressive search, vehicle re-identification, deep learning, license plate verification, contextual information.**

I. Introduction

VEHICLES, such as cars, buses, and trucks, have been an indispensable part of human life as well as an important class of objects in urban surveillance systems. Many researchers in the multimedia and computer vision fields have focused on vehicle-related research, such as detection [1], fine-grained categorization [2], 3-D pose estimation [3], and driver behavior modeling [4]. Nevertheless, **vehicle re-identification (Re-Id)** is a significant but frontier area that is always overlooked and far from solved by the community. Taking a query vehicle as the input, vehicle Re-Id aims to search in the surveillance data and find the same vehicle recorded by different cameras, as shown in Fig. 1. Vehicle Re-Id can be pervasively applied in intelligent surveillance systems [5], smart transportation [6], and urban computing [7]. Through the ubiquitous surveillance networks, it can quickly tell users where and when the vehicle was in the city.

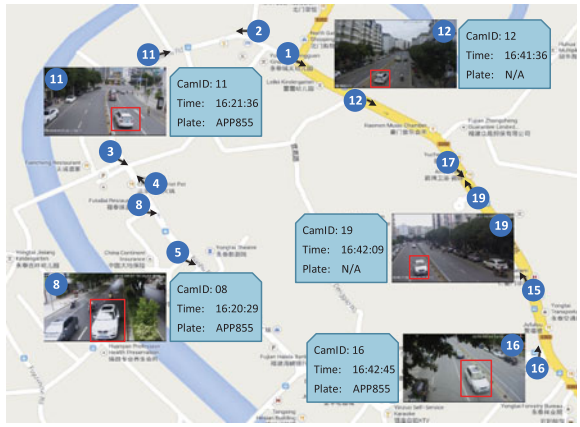

Fig. 1. An example of vehicle Re-Id: searching for the white BMW SUV with multi-modality data in urban surveillance.

Vehicle Re-Id can be considered as an instance-level object search task, which is different from traditional vehicle detection, tracking, and categorization problems. Similar to near-duplicate image retrieval [8], [9], content-based video search [10], and object instance search [11], vehicle Re-Id is to find the vehicle with the same identity from urban surveillance videos. In real-world practice, humans can treat this task in a progressive manner [12]. For instance, if the security officers need to find a suspect car in a city with large-scale video surveillance networks, appearance attributes such as models, types, and colors can be initially used to find similar vehicles and reduce the search field. Then, they can identify the targets precisely from the filtered vehicles by matching the license plates, which can reduce the enormous workload. Meanwhile, they will search videos recorded by cameras from near to far positions and from close to distant time range. Therefore, the contextual information

Manuscript received January 26, 2017; revised May 7, 2017 and June 27, 2017; accepted August 6, 2017. Date of publication September 13, 2017; date of current version February 14, 2018. This work was supported in part by the National Key Research and Development Plan under Grant 2016YFC0801005, in part by the Funds for Creative Research Groups of China under Grant 61421061, in part by the National Natural Science Foundation of China under Grant 61602049, and in part by the Beijing Training Project for the Leading Talents in S&T under Grant ljrc 201502. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shu-Ching Chen. *(Corresponding author: Huadong Ma.)*

X. Liu, W. Liu, and H. Ma are with the Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: xinchenliu@bupt.edu.cn; liuwu@bupt.edu.cn; mhd@bupt.edu.cn).

T. Mei is with the Microsoft Research Asia, Beijing 100080, China (e-mail: tmei@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2751966

(a) (b) (c)

Fig. 2. (a) The same vehicles have great within-class differences in different viewpoints (left). Different but similar vehicles have trivial inter-class differences (right). (b) The license number plates as the unique ID for vehicle search. (Plate is masked to protect privacy.) (c) The contextual information can assist in vehicle search in the city.

such as spatiotemporal cues thus can decidedly assist in the search process. Inspired by real-world practice, we can construct a progressive vehicle search framework in a two-step procedure with multi-level attributes and multi-modal data: 1) searching from coarse to fine in the feature domain, which first employs the appearance features for a coarse but fast filter and then exploit the license plate as the unique identifier to find the same vehicles; and 2) searching from close to far in the physical world, which considers the time and locations as the key cues for vehicle search.

Nevertheless, the construction of the progressive vehicle re-identification framework with multi-modal data from practical urban video surveillance faces three significant challenges: first the appearance-based methods usually cannot give satisfactory results because of the trivial inter-class differences between different vehicles from similar viewpoints and the dramatic within-class differences of the same vehicle from various viewpoints, as shown in Fig. 2(a). Moreover, conventional license plate recognition systems can hardly recognize the license plate in an unconstrained surveillance environment because of the various lightning conditions and viewpoints, noise, and low resolution, as shown in Fig. 2(b). In addition, the plate recognition system usually contains multiple procedures such as plate localization, calibration, character segmentation, and recognition, as in [13], [14]. If one of the steps fails or any of the characters on the plate is mis-recognized, the vehicle Re-Id results might be incorrect. How to utilize the license plate effectively and efficiently in unconstrained urban surveillance is a crucial challenge. Furthermore, the contextual information, such as the spatiotemporal pattern of vehicles, camera locations, and topology of the city roads is difficult to discover and model. The environmental factors and the driver's behavior can introduce great uncertainty [4]. How to utilize the contextual information is another great challenge.

Existing vehicle Re-Id approaches are predominantly focused on appearance features of vehicles, such as colors, types, shapes, and detailed attributes [1], [15]–[17]. Therefore, they can hardly differentiate among vehicles with similar models and colors and identify the same vehicle in a varied environment. Moreover, they usually overlook unique identifiers, such as number plate when matching a vehicle. In contrast, we comprehensively utilize the appearance attributes and the license plate information in a coarse-to-fine manner for vehicle search. The appearance features can be employed to find the similar vehicles, and then the license number plates are used to match the same vehicle precisely. In addition, existing approaches neglect the spatiotemporal context. Contextual information has been exploited in several research fields such as intelligent surveillance [18], cross-camera person tracking [19], person Re-Id [20], and object retrieval [21]. With contextual cues recorded by the surveillance system, we treat the search procedure by a from-close-to-far manner in the physical space.

This paper proposes a PROgressive Vehicle re-IDentification framework based on deep neural networks, named PROVID, which features four important properties: 1) a progressive vehicle Re-Id paradigm is designed to exploit multi-modality data in urban surveillance such as multi-level visual features, license plates, camera locations, and contextual information; 2) the appearance of the target vehicle is used as a coarse filter by integrating hand-crafted features and high-level attributes learned by convolutional neural network; 3) a Siamese neural network is adopted to verify license number plates for precise vehicle search; and 4) a spatiotemporal model is exploited to further improve the search procedure. Particularly, we consider the plates as the fingerprints of vehicles, and we just need to verify two plate images instead of precisely recognizing the characters. Furthermore, a spatiotemporal relation (STR) model is designed as the context to re-rank the results.

To evaluate the proposed framework and facilitate related research, "VeRi", a comprehensive vehicle Re-Id dataset, is constructed from a practical urban video surveillance system. It includes not only large numbers of vehicles with various annotations and sufficient cross-camera recurrences but also plenty of license plates and spatiotemporal information. Extensive experiments on the VeRi dataset demonstrate that our PROVID framework achieves excellent accuracy and speed. Finally, we discuss several extension of the progressive search, which can be utilized in various applications.

Compared with our previous works [15], [22], we propose a Null space based Fusion of Color and Attribute feaTure model (NuFACT), which can significantly improve the accuracy for appearance-based vehicle search, e.g., 29.73% in mean Average Precision (mAP) and 24.55% in HIT@1. In [15], [22], the texture, color, and high-level attributes are fused by direct

early-fusion or late-fusion strategy, while the NuFACT adopts a Null Foley-Sammon Transform (NFST)-based metric learning approach for fusion of multi-level features. It can not only learn discriminative representation of vehicle appearance from different viewpoints but also reduce the feature redundancy (from approximately 7,000-D to 1,000-D) to guarantee efficiency. To evaluate the adaptation ability of PROVID under different conditions, we conduct extensive experiments on two large-scale vehicle Re-Id datasets, i.e., VeRi [22] and VehicleID [16]. Comprehensive experiments demonstrate that PROVID not only dramatically improves the accuracy but also reduces the computational cost for vehicle Re-Id.

II. Related Work

A. Vehicle Re-Identification/Search

Vehicle search, or Re-Id, is a frontier area with limited related research in recent years. Feris *et al.* [1] designed a vehicle detection and retrieval framework. They first classified vehicles by type, size, and color, and then organized and retrieved vehicles with a relational database. Yang *et al.* [2] proposed the adoption of the deep convolutional neural network for fine-grained vehicle categorization, model verification, and attribute prediction, and collected a vehicle image dataset, CompCars, to validate the proposed method. Recently, Liu *et al.* [15] explored some appearance features, such as the texture, color, and semantic attributes learned by convolutional neural networks. They also built an appearance-based model by integrating low-level and high-level semantic features for vehicle search. Liu *et al.* [16] proposed a Deep Relative Distance Learning (DRDL) framework, which could jointly learn the feature representation and metric mapping. Nevertheless, appearance-based methods can hardly distinguish among similar vehicles from the same viewpoints and identify the same vehicle under different conditions, such as various illuminations and viewpoints. Additionally, the license plate, as the distinct property of vehicles, should be utilized to precisely identify the same vehicle. Furthermore, existing datasets, such as CompCars [2] and VehicleID [16], only provide the appearance labels such as types and models, neglecting the license plate and contextual information, which are important for vehicle Re-Id in large-scale urban surveillance.

B. License Plate for Vehicle Search

In real-world practice, parks and highways have adopted license plate recognition systems to identify vehicles [13], [14]. However, existing systems require high-quality license plate images. Therefore, the cameras are usually installed in constrained situations such as entrances of parks or toll gates of highways, calibrated with proper viewpoints, and require auxiliary infrastructure such as flashlights and sensors. While in unconstrained traffic environments, the license plate recognition system can not work well because of uncertain factors such as various lightning conditions and occlusions [1], [15]. Thus, we propose to verify the license plates instead of recognizing all characters of the plates. Recently, deep learning models, such as convolutional neural networks (CNNs), have obtained state-of-the-art results in many multimedia and vision tasks such as image categorization [23], object detection [24], image analysis [25], video summarization [26], and multimedia retrieval [27]. In particular, Bromley *et al.* [28] proposed a Siamese Neural Network (SNN) for hand-written signature verification. SNN is built with two CNNs with shared parameters to extract discriminative features, and trained by the contrastive loss to learn a latent space for the similarity metric. Chopra *et al.* [29] employed the SNN to verify faces and achieved state-of-the-art results. Zhang *et al.* [30] propose to identify persons with gait features learned by SNN and obtain significant improvement. Inspired by these methods, we adopt SNN to verify license plate in our vehicle Re-Id framework.

C. Contextual Models

Contextual information, e.g., the spatiotemporal records, object locations, and topology of cameras, has been widely exploited in multi-camera systems [18], [19], [21]. For examples, Kettnaker *et al.* [18] adopted a Bayesian estimation model to assemble likely paths of objects over different cameras. Javed *et al.* [19] proposed to estimate the inter-camera correspondence with spatiotemporal information for cross-camera person tracking. Recently, Xu *et al.* [21] designed a graph-based object retrieval framework to find persons and cyclists on the campus. However, existing approaches usually consider objects that move at low speed, such as persons and cyclists. In addition, they mainly focus on constrained environments, e.g., parks, campuses, and buildings. In an urban area, the traffic scenes, such as roads and crossroads, are mostly unconstrained environment with significant uncertainty due to the complex environments and varied road topology. We can still gain some insights from the above works to exploit the contextual cues for vehicle Re-Id.

III. Overview of the PROVID Framework

In Fig. 3, we show the architecture of the PROVID framework. In our framework, the input query is a vehicle image and contextual information from the surveillance system, e.g., the camera ID and spatiotemporal cues. With the query, the PROVID framework can search for the same vehicle by three procedures: 1) coarse filtering by vehicle appearance: the framework utilizes the appearance model to find the vehicles that have similar texture, shape, color, and type in surveillance videos; 2) precise search by license plate verification: with the Siamese neural network, the license plate distances between the query vehicle and gallery vehicles are estimated for the filtered vehicles to match the same vehicles; 3) the spatiotemporal relation model (STR) is proposed to re-rank the previous results and identify the optimal vehicles.

IV. Vehicle Filtering by Appearance

A. Multi-Level Vehicle Representation

In practical vehicle search, it is effective to filter vehicles by appearance features, e.g., texture, shape, type, and color. Besides, these features can be extracted and matched efficiently in large-scale data.

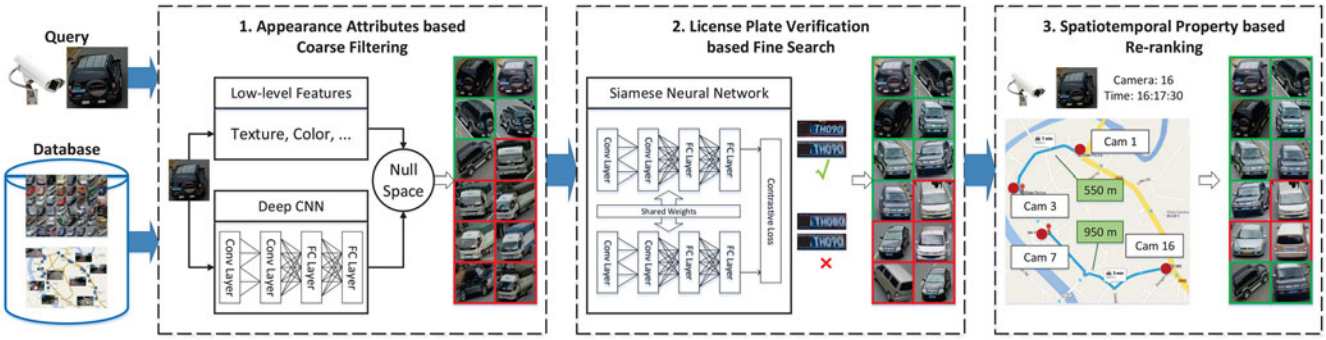

Fig. 3. The architecture of the PROVID framework.

In our previous work [15], we propose to use multi-level appearance feature as the coarse filter to search for the vehicles that have similar appearance. For the **texture feature**, we adopt the traditional Scale-Invariant Feature Transform (SIFT) [31] as the local descriptor. Then, the bag-of-words (BOW) model is used to quantized the SIFT descriptor because of the efficiency and effectiveness in multimedia retrieval [32]. For the **color feature**, the Color Name (CN) descriptor [33] is extracted and then encoded by the BOW for high-accuracy person re-identification [34]. For the **high-level semantic features**, we exploit a deep convolutional neural network (CNN), i.e., the GoogLeNet [35], as the feature extractor. The CNN is pre-trained on the ImageNet dataset [36] and fine-tuned on the CompCars dataset [2] which has been labeled with many detailed attributes, e.g., the light shape, the number of seats, the number of doors, and the vehicle model. Therefore, by fine-tuning on CompCars, the model can learn many rich high-level semantic features that are very effective for vehicle search.

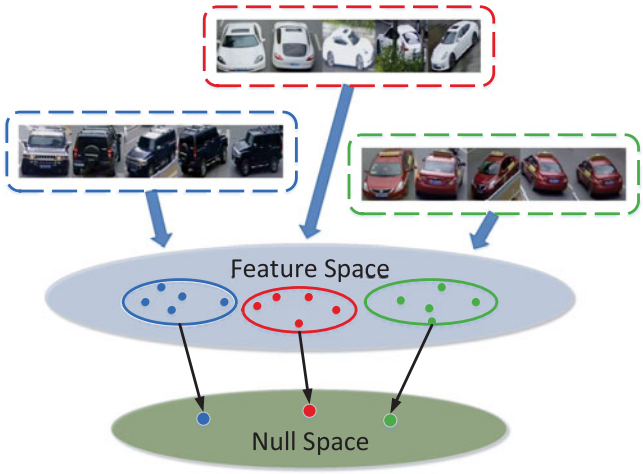

Fig. 4. The appearance features of the same vehicle are mapped to a single point by NFST.

B. The Null-Space-Based FACT Model

The FACT model in [15] adopted a post-fusion scheme to directly sum the Euclidean distances of three types of features extracted from vehicle images. However, it cannot effectively integrate the complementary multi-level features. The Null Foley-Sammon Transform (NFST) was first proposed to address the small sample size problem in face recognition [37]. Zhang *et al.* [38] proposed a Kernelized NFST for person Re-Id by mapping the multiple features into a discriminative null space; this method significantly outperforms the state-of-the-art methods. In this paper, we propose a Null-space-based FACT (NuFACT) to extract effective and robust representations for vehicle appearance.

The NFST is one type of metric learning methods; other examples of metric learning methods include Linear Discriminant Analysis (LDA) and Foley-Sammon Transform (FST) [39]. The basic idea of the FST is to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$ and maximize the Fisher discriminant criterion:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}, \tag{1}$$

where $\mathbf{w}$ denotes a column of $\mathbf{W}$, and $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between-object scatter matrix and within-object scatter matrix, respectively. With $\mathbf{W}$, the original visual features can be mapped into a latent metric space in which the distances of features from the same object are much smaller than those of features from different objects.

However, NFST aims to learn a null space by adopting an extreme restrictive constraint:

$$\mathbf{w}^\top \mathbf{S}_w \mathbf{w} = 0, \tag{2}$$

$$\mathbf{w}^\top \mathbf{S}_b \mathbf{w} > 0. \tag{3}$$

In the null space, the features of each object are collapsed to a single point, which means the intra-object distance is zero and inter-class distance is positive, as shown in Fig. 4.

Furthermore, to learn a discriminative null space for person Re-Id [38], Zhang *et al.* introduce a kernel function $\Phi(\mathbf{x})$ to NFST that can map the original feature $\mathbf{x}$ into an implicit high-dimensional space. During learning of the discriminative null space on the training data, the multiple features are fused effectively and can generate a discriminative representation for person Re-Id.

In this paper, we adopt the discriminative NFST method to integrate the multi-level features of vehicles, i.e., texture, color, and high-level attribute features. First, the three types of features, $\mathbf{X}_t$, $\mathbf{X}_c$, and $\mathbf{X}_a$, are extracted from all training vehicle images and concatenated to obtain the original appearance

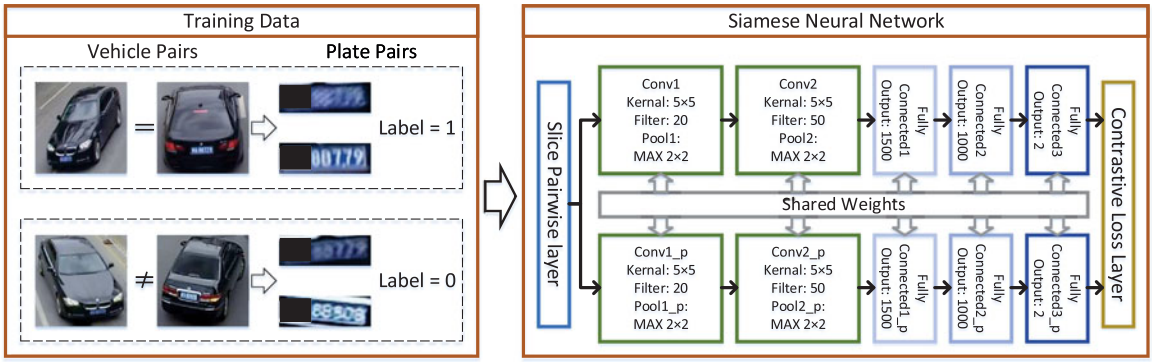

Fig. 5. The architecture of license plate verification based on the Siamese neural network.

feature as $\mathbf{X} = (\mathbf{X}_t, \mathbf{X}_c, \mathbf{X}_a)$. Then, the training features $\mathbf{X}$ are kernelized by $\Phi(\mathbf{x})$ to obtain $\Phi(\mathbf{X})$. Finally, the projection matrix $\mathbf{W}$ of the discriminative null space is learned by NFST on $\Phi(\mathbf{X})$ as in [38].

In the test phase, the original features $\mathbf{X}_q$ and $\mathbf{X}_g$ of the query and gallery vehicles are also kernelized with $\Phi(\mathbf{x})$ and mapped by $\mathbf{W}$. Finally, the similarity of the query and gallery vehicles can be measured by the Euclidean distance in the discriminative null space. By NFST-based multi-level feature fusion, the vehicles that have the similar appearance to the query are obtained effectively and efficiently. After this procedure, a small number of vehicles are extracted from the whole database of vehicles. Nevertheless, it can hardly uniquely match images of the same vehicle based only on appearance features, which cannot distinguish similar vehicles with trivial inter-class differences due to environmental factors. In these situations, the distinct identifier, i.e., the license plate, must be considered for precise vehicle search.

V. License Plate Verification Based on Siamese Neural Network

As shown in Fig. 2(b), the characters on a license plate can hardly be recognized correctly in unconstrained environments because the varied viewpoints and lightning conditions cause the plate images to be blurry. In addition, license plate recognition systems are usually composed of several components such as plate detection, calibration, character segmentation, and recognition. Thus, the license plate recognition techniques are unsuitable for the vehicle Re-Id task. Therefore, we propose to verify the license plate instead of recognizing the plate number for precise vehicle search. The Siamese neural network (SNN) proposed by Bromley *et al.* [28] was originally designed to verify hand-written signatures. SNN is built with convolutional layers to discover the feature representation and fully-connected layers to learn a mapping function from the large number of training images. With SNN, the discriminative features can be extracted directly from image pairs, and then the features are mapped into a metric space in which the distance between different objects is large while the distance between the same objects is small. Therefore, SNN is very suitable for tasks in which there are large numbers of objects but the samples of all the classes are insufficient. Decidedly, SNN can be adopted for license plate verification which has this property.

In our framework, we designed the SNN for plate verification as illustrated in Fig. 5. Two parallel CNNs have the same structure and share the same weights in forward and backward computations. Each CNN is built with two convolutional layers and max-pooling layers for feature representation, and three fully connected layers to learn the metric space. The detailed parameters are selected as shown in Fig. 5. In the training phase, a pair of license plate images is assigned a value 1 if they have the same number and 0 otherwise. After that, the contrastive loss layer takes the output features of the last layer and the labels as the input to calculate the cost of the model. With the Stochastic Gradient Descent algorithm, the SNN is optimized with the contrastive loss.

In particular, we denote by W the weights of the neural network, and by x_1 and x_2 a pair of input plates. The features obtained by the forward propagation can be denoted by $S_W(x_1)$ and $S_W(x_2)$. The difference between x_1 and x_2 is denoted as

$$E_W(x_1, x_2) = ||S_W(x_1) - S_W(x_2)||. \tag{4}$$

With $E_W(x_1, x_2)$, the contrastive loss is defined as

$$L(W, (x_1, x_2, y)) = (1-y) \cdot max(m - E_W(x_1, x_2), 0) + y \cdot E_W(x_1, x_2), \tag{5}$$

where (x_1, x_2, y) is a three-tuple of two training plates and the corresponding label, and m is a positive hyperparameter to adjust the margin ($m = 1$ in our method). In our framework, the Caffe deep learning tool [40] is adopted to implement the SNN and train the model. In the testing phase, the output of the second fully connected layer (FC2) in the learned SNN is extracted as the 1,000-D feature representation for the plate images. Finally, the similarity of two input plates is computed by the Euclidean distance.

VI. Spatiotemporal Relation-Based Vehicle Re-Ranking

In practical vehicle search, humans usually execute the search process in a close-to-far manner in the physical world. Therefore, the spatiotemporal information is explored in our progressive vehicle Re-Id framework. Nevertheless, how to model the

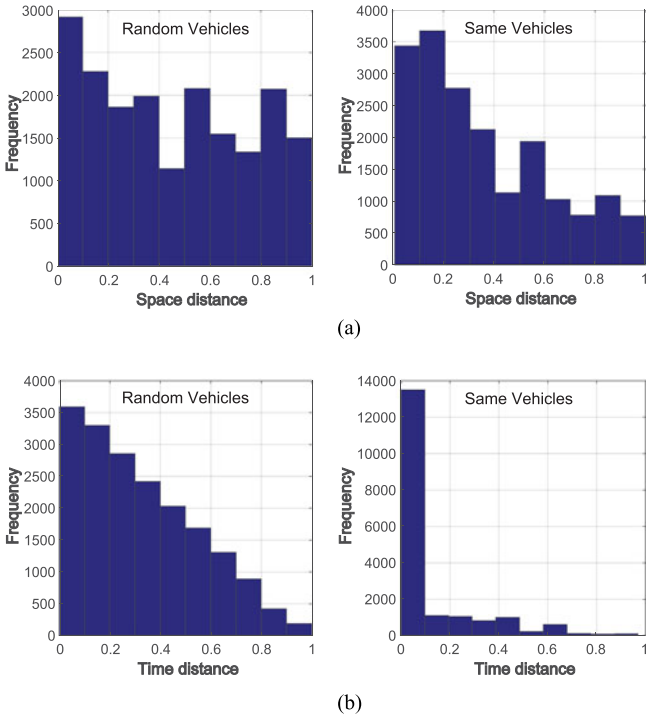

Fig. 6. Statistics of spatiotemporal information. (a) Histograms of space distances. (b) Histograms of time distances.

behavioral features of vehicles and discover the spatiotemporal property of the same vehicle remains a significant challenge, especially in unconstrained environments and with only video surveillance networks.

To explore the effect of spatiotemporal information for vehicle Re-Id in unconstrained scenes, we select 20,000 pairs of the same vehicles and 20,000 pairs of vehicles that are picked randomly. Then, the spatiotemporal difference of each pair is calculated for analysis. The histograms in Fig. 6 show the statistics (the spacial distances and temporal distances of all samples are normalized to $[0, 1]$ for better representation). It is obvious that the pairs of the same vehicles have smaller spatiotemporal differences than the pairs of randomly selected vehicles. Hence, an assumption is made based on this observation: two images are more likely to be the same vehicle when their spatiotemporal difference is small, whereas they are more likely to be the different vehicles when their spatiotemporal difference is large. Based on this assumption, given a pair of images i and j, $ST(i, j)$ is the spatiotemporal similarity formulated as:

$$ST(i,j) = \frac{|T_i - T_j|}{T_{max}} \times \frac{\delta(C_i, C_j)}{D_{max}} \tag{6}$$

where T_i and T_j are the timestamps at which the images are captured by the cameras and T_{max} is a global maximum value obtained from all vehicle images captured over a long time period. $\delta(C_i, C_j)$ is the physical distance between camera C_i and C_j, and D_{max} is a global maximum distance between all cameras. The physical distance between each pair of cameras is obtained from a public online map services, i.e., Google Maps, and organized as a distance matrix as illustrated in Fig. 7. In our framework, we assume the distance matrix is symmetric which

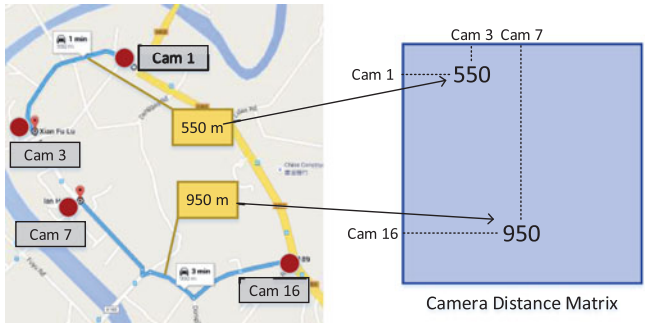

Fig. 7. The physical distance matrix of surveillance cameras.

means the distances from camera C_i to C_j and from camera C_j and C_i are equal. Finally, the spatiotemporal similarity can be integrated with the appearance and plate features using the late fusion or the top-K re-ranking scheme for efficiency.

VII. Applications

In this section, we show how the progressive vehicle Re-Id framework can be utilized in various practical applications.

A. Application I: Suspect Vehicle Search

As its core functionality, our PROVID framework can support suspect vehicle search for vehicle and traffic management department. Consequently, with a query vehicle image captured by a surveillance camera, users can instantly obtain information on where and when the vehicle has ever appeared in the whole city. Our framework can incorporate cameras deployed in constrained scenes in which the license plate recognition system can be applied. Then, the vehicles can be searched more accurately with cameras in both constrained and unconstrained environments. With more precise license plate information, detailed information about the vehicle can be found by users. For example, our PROVID system can be integrated with cameras at the park entrances or toll gates, and then connected with the vehicle registration information system. When security officers have an image of a suspect car, they can first use our system to find the locations and time at which the car appeared. Then, they can use the license plate recognition system to obtain its license number via the toll gate camera. With the license number, detailed information such as the owner of the vehicle, registration time, and criminal records can be searched from the vehicle database in the registration system. Using this information, the staff can manage the vehicles or investigate criminal events more effectively and efficiently. In summary, our progressive vehicle Re-Id system becomes a vehicle search engine for urban surveillance networks.

B. Application II: Cross-Camera Vehicle Tracking

The proposed vehicle search framework can also be applied to track the target vehicle across multiple cameras. For example, if the police officers want to track a suspect car in the city, they can first specify a target vehicle in one camera from the backend browser. Then, our progressive vehicle search system can take

the vehicle image, location, and time as input to find the same vehicle in the neighboring cameras. Consequently, the system can track the target vehicle from one camera to another and obtain the route of the target. It can provide significant assistance for criminal investigation and urban security. Another example is live broadcasts of car races. The car races such as Dakar Rally or Formula One are usually broadcasted by multiple cameras. In particular, viewers are willing to watch a specific car in videos from different cameras at different time while all cars look very similar. With the vehicle search system, the users or directors can specify the car that needs to be tracked at a specific time. Then, the system can instantly track the target car by the appearance and unique identifiers, such as the numbers or names on the car. In conclusion, our system can help users localize and track vehicles across multiple cameras automatically, which is very useful for suspect car tracking in urban surveillance and live broadcasts of car races.

VIII. Experiments

A. Dataset

1) VeRi Dataset: To facilitate related research and evaluate the proposed progressive vehicle search framework, we build a comprehensive vehicle Re-Id dataset, named VeRi. A total of 20 surveillance cameras installed along several roads in a 1.0 km^2 area are selected to guarantee data quality and real-world traffic scenarios. Various scenes are captured by the cameras, such as crossroads, two-lane roads, and four-lane roads. The cameras record videos at a resolution of 1920 $\times$ 1080 and 25 frames per second. The cameras are installed in arbitrary positions and directions (the orientation and tilt-angle information is not available). In addition, overlaps exist between part of the cameras. The construction process of the VeRi dataset is introduced in our previous papers [15], [22]. Fig. 8 shows some sample images and main statistics of the dataset.[1]

The VeRi dataset has four featured properties that make it a valuable and challenging dataset:

1) *Large-scale data from real-world surveillance*: We select continuous one-day raw videos from 20 surveillance cameras. Then, the videos from 16:00 to 17:00 are segmented from the original videos with basic compression and transcoding. To balance quality and efficiency, one in every five frames is extracted from the 25-fps videos to obtain over 360,000 frames for vehicle annotation. After the annotation in [15], we obtain approximately 50,000 images and 9000 tracks of 776 vehicles, which guarantee the scalability for vehicle search. Each vehicle is captured by at least two cameras from various viewpoints, lightning conditions, and backgrounds which guarantees a practical urban traffic environment, as shown in Fig. 8(a), and sufficient cross-camera recurrence for vehicle search, as shown in Fig. 8(b). The dataset is split into a training set containing 37,781 images of 576 vehicles and a testing set with 11,579 images of 200 vehicles. From the testing set, we select one image from each camera and of each vehicle as the query and obtain a query set containing 1,678 images.

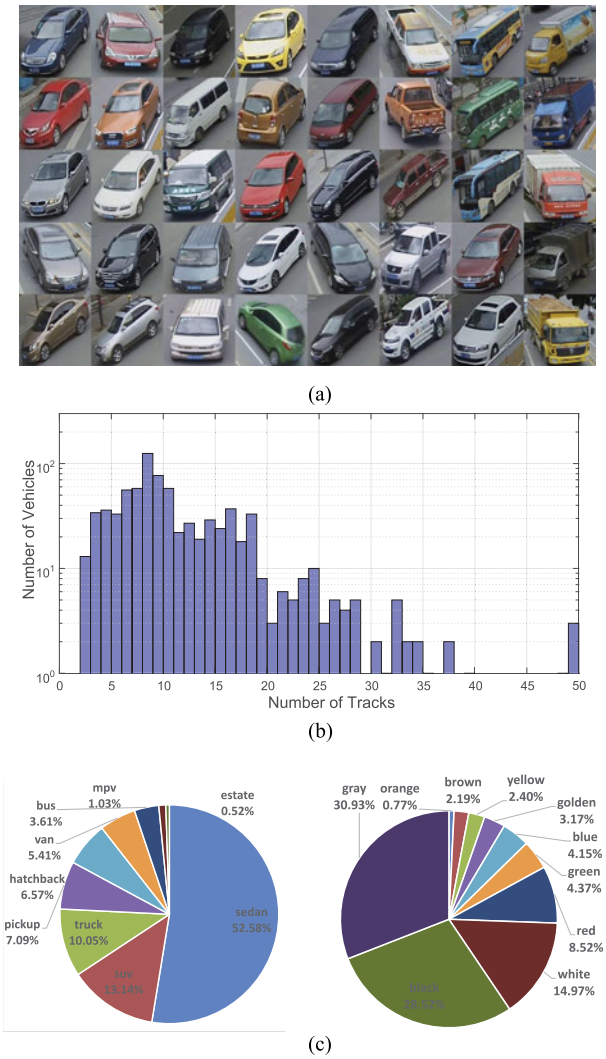

Fig. 8. The main properties of the VeRi dataset.(a) Sample images in VeRi dataset. (b) Distribution of numbers of vehicle tracks. (c) Statistics of types and colors.

2) *Rich attribute labels*: Each vehicle image in the VeRi dataset is labeled with various attributes. First, we annotate the bounding boxes (BBoxes) as well as the locations of the vehicle images in video frames which can also be used for vehicle detection tasks. Moreover, we annotate 10 types of colors, i.e. black, gray, white, red, green, orange, yellow, golden, brown, and blue to label the color of vehicles. Furthermore, each vehicle is labeled with one of nine classes, i.e., sedan, SUV, hatchback, MPV, van, pickup, bus, truck, and estate car. In addition, part of the vehicles are labeled with about 30 common brands, such as BMW, Audi, Ford, and Toyota. The statistics of colors and types are shown in Fig. 8(c).

3) *License plate annotation*: As one of the most noteworthy contribution of the VeRi dataset, we annotate the license number plate if it can be detected in the vehicle image by

[1]The latest version of the VeRi dataset can be obtained from https://github.com/VehicleReId/VeRidataset

the annotators. For each image in the training, testing, and query sets, we annotate the location of the license plate and the characters if they can be recognized. At least three annotators are asked to label each image to guarantee high quality. Finally, 999, 4,825, and 7,647 plates are obtained from the query, testing and training sets respectively.

4) *Contextual information annotation*: As important contextual information, the spatiotemporal information of vehicles, camera topology, and distances between cameras are annotated. Firstly, we annotate the camera ID which records the vehicle track and the time at which it is captured. Then, the distance between each pair of cameras in the surveillance system is obtained from Google Maps, as shown in Fig. 7. With the above contextual information, the multi-modal data can be exploited for progressive vehicle Re-Id.

2) VehicleID Dataset: Recently, Liu *et al.* [16] built a large-scale dataset for vehicle re-identification named VehicleID. It contains images captured in the daytime by different cameras in the traffic surveillance system of a small city. Similar to our VeRi dataset, each vehicle appears more than one time in different cameras. It contains a total of 26,267 vehicles with 221,763 images, and 10,319 vehicles are labeled with models such as Ford Focus, Toyota Corolla, and Honda Accord. To facilitate the research, the VehicleID dataset is split into a training set with 110,178 images of 13,134 vehicles and a testing set with 111,585 images of 13,133 vehicles. In addition, from the original testing data, three subsets, which contain 800, 1600, and 2400 vehicles, are extracted for vehicle search in different scales.

There are two main differences between our VeRi dataset and the VehicleID dataset. First, although the scale of VehicleID is larger than VeRi, the vehicles of VehicleID are captured only from the front or the back, whereas our dataset contains vehicle images captured by 20 cameras with various viewpoints, resolutions, and occlusions, which can reflect practical situations. This makes VeRi closer to a real-world unconstrained environment and more challenging for vehicle Re-Id. Furthermore, VehicleID can only be used for appearance-based vehicle Re-Id or related research. In addition to vehicle images, our dataset contains license plate annotations and spatiotemporal information. This means that VeRi can not only facilitate vehicle Re-Id in a surveillance network but also provide potential value for license plate recognition, traffic data mining, and urban computing.

B. Experimental Settings

In this paper, we first compare different appearance-based method on both of the VehicleID and VeRi dataset. Then, we evaluate the license-plate-based vehicle search and the complete progressive PROVID framework on the VeRi dataset.

For VehicleID, image-to-image search is conducted because each vehicle is captured in one image by one camera. For each test dataset (size = 800, 1600, and 2400), one image of each vehicle is randomly selected into the gallery set. All other images are probe queries. To measure the accuracy of the approaches, we adopt HIT@1, HIT@5, and Cumulative Matching Characteristic (CMC) curve, as in [16].

For VeRi, cross-camera matching is performed, which means that one image of a vehicle from one camera is used as the query to search images from other cameras for the same vehicle. In addition to the image-to-image search as for VehicleID dataset, we also adopt an image-to-track approach, in which the image is used as the query, while the gallery consists of tracks of the same vehicle captured by other cameras. A track is a trajectory of a vehicle recorded by one camera at a time, which means the images in a track are organized together. The similarity between an image and a track is computed by max-pooling over images in the test track because, in the practical search procedure of humans, it is reasonable to find the most possible image in the track from one camera to capture the target vehicle. Therefore, we use 1,678 query images and 2,021 testing tracks for the image-to-track search. The CMC curve, HIT@1 (precision at rank 1), and HIT@5 (precision at rank 5) are also adopted to evaluate the accuracy of the methods. In addition, the query has more than one ground truth, so precision and recall should be considered in our experiments. Hence, we also use mean average precision to evaluate the comprehensive performance. The average precision (AP) is computed for each query as

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{gt}} \tag{7}$$

where n and N_{gt} are the numbers of tests and ground truths respectively, $P(k)$ is the precision at the k-th position of the results, and $gt(k)$ is an indicator function that equals to 1 if the kth result is correctly matched and 0 otherwise. Over all queries, the mean Average Precision (mAP) is formulated as

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{8}$$

in which Q is the number of queries.

C. Evaluation of Appearance-based Vehicle Re-Id

In this experiment, we compare eight vehicle Re-Id approaches which are evaluated on both VehicleID and VeRi. The details of the approaches are introduced as follows.

1) *Texture based feature (BOW-SIFT)*: For both VeRi and VehicleID datasets, the image is resized to 64×128 firstly. Then, we extract the SIFT local descriptors [31] from the images. After that, the descriptors are encoded by the BOW model with the pre-trained codebook (size $k = 10,000$). Finally, we obtain a 10,000-D feature to represent the texture of the vehicle.
2) *Local Maximal Occurrence Representation (LOMO)*: LOMO is proposed as a local feature for person Re-Id that is robust to the varied lightning conditions in practical surveillance scenes [41]. We consider LOMO as the state-of-the-art texture feature. For both the VehicleID and VeRi, we extract the LOMO feature with the parameters given in [41] and obtain a 26,960-D feature vector for each vehicle image.
3) *Color based feature (BOW-CN)*: This model is the benchmark for person Re-Id on the Market-1501 dataset [34] due to its robustness in outdoor scenes. It first adopts the

Color Name (CN) [33] as a local color descriptor. Similar to BOW-SIFT, the image is resized to 64×128. Then, we divide the image into 4×4 patches to extract the CN descriptors densely. Before testing, a pre-trained codebook is built on VeRi and VehicleID separately using k-means (size $k = 350$). After that, the avgIDF and geometrical priors are applied as in [34]. Finally, a 5,600-D color feature is obtained for each image.

4) *Semantic feature learned by CNN (GoogLeNet)*: For VeRi, we adopt the GoogLeNet model [35] pre-trained on ImageNet [36]. As in [2], the model is fine-tuned on the CompCars dataset, which contains images of whole and parts of cars with rich attributes such as the number of doors, the light shape, and the car model. The finetuned CNN model is employed as a feature extractor for high-level attributes. Finally, we obtain a 1,024-D feature from the last pooling layer of the neural network to represent the semantic feature of vehicles.
5) *Fusion of Attributes and Color feaTures (FACT)*: As in [15], by combining the low-level color feature and high-level semantic attribute, the FACT model achieves excellent performance on the VeRi dataset. We implement the FACT model on both VeRi and VehicleID. The fusion weights are obtained on a small subset of the training data for validation.
6) *Deep Relative Distance Learning with VGG (DRDL-VGG)*: The DRDL framework is proposed to jointly learn a discriminative feature representation and a metric mapping with an end-to-end CNN and achieves the state-of-the-art results on the VehicleID dataset [16]. It adopts a mixed network structure based on the VGG_M model [42] with a coupled cluster loss to learn the relative distances of different vehicles. Because the VeRi dataset does not contain model information as VehicleID, we only evaluate DRDL-VGG on VehicleID.
7) *Semantic feature learned by VGG (VGG)*: To evaluate different deep-learning-based models, we directly use the VGG_M model in DRDL-VGG [16] as a feature extractor for testing on the VeRi dataset. The 1024-D feature is extracted from the fc_7 layer of the VGG_M model.
8) *Null space base Fusion of Attribute and Color feaTures (NuFACT)*: As introduced in Section IV-B, we concatenate the color feature and the semantic attributes to obtain the original features of vehicles for VeRi and VehicleID separately. Then, the projection matrix to the null space is learned on the corresponding training sets. Finally, we evaluate the NuFACT model on both VeRi and VehicleID.

Table I illustrates mAP, HIT@1, and HIT@5 on VeRi, Fig. 9 shows the CMC curves. The results on VehicleID are shown in Table II and Fig. 10. From the results, we obtain the following findings:

1) For both VehicleID and VeRi datasets, the hand-crafted features, i.e., BOW-SIFT, LOMO, and BOW-CN achieves relatively lower accuracy than the deep learning-based models, i.e., GoogLeNet and VGG. This demonstrates that the features learned by deep neural networks are more discriminative and robust than conventional features for vehicle Re-Id. Moreover, the fusion model of multi-level features, i.e., FACT, and the mixed neural network structure, i.e., DRDL, obtain higher accuracy than the above single-model approaches. This shows that the high-level attributes and low-level hand-crafted features have complementary effects for vehicle Re-Id. Finally, our proposed NuFACT model achieves the optimal results on both VehicleID and VeRi. This means that, in the null space learned by NFST, the multi-level features can be fused effectively for vehicle Re-Id.
2) By comparison of the results on the two datasets, we find that different methods have different characteristics. First, the texture feature, i.e., LOMO, has better accuracy than the color feature on VehicleID, while we obtain opposite results on VeRi. By the examining the two datasets, we find that the vehicle images in VehicleID are relatively larger and sharper than the images in VeRi. More detailed texture can be extracted from the images in VehicleID than in VeRi for the LOMO. Besides, some of the images in VehicleID are captured at night and are almost black in hue, while VeRi contains only images captured in the daytime. Therefore, we can obtain more effective color features from the images in VeRi than from those in VehicleID. Second, NuFACT achieves much better improvement on VeRi than on VehicleID. One reason is that the color feature is more effective on VeRi than on VehicleID, so the

TABLE I
THE IMAGE-TO-TRACK SEARCH RESULTS ON VERI

methods	mAP	HIT@1	HIT@5
BOW-SIFT	1.51	1.91	4.53
LOMO [41]	9.64	25.33	46.48
BOW-CN [34]	12.20	33.91	53.69
VGG [16]	12.76	44.10	62.63
GoogLeNet [2]	17.89	52.32	72.17
FACT [15]	18.75	52.21	72.88
NuFACT	**48.47**	**76.76**	**91.42**

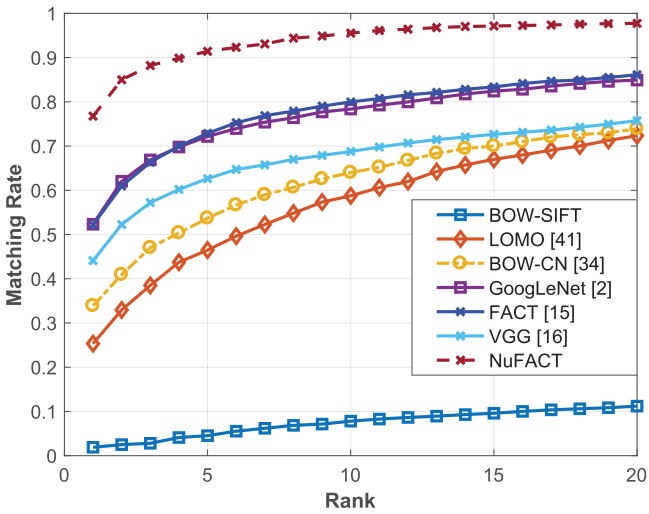

Fig. 9. The CMC curves of different methods on VeRi.

TABLE II
COMPARISON OF DIFFERENT METHODS ON VEHICLEID

Methods	Test size = 800		Test size = 1600		Test size = 2400		Average	
	HIT@1	HIT@5	HIT@1	HIT@5	HIT@1	HIT@5	HIT@1	HIT@5
BOW-SIFT	2.81	4.23	3.11	5.22	2.11	3.76	2.68	3.76
LOMO [41]	19.74	32.14	18.95	29.46	15.26	25.63	17.98	3.76
BOW-CN [34]	13.14	22.69	12.94	21.09	10.20	17.89	12.09	20.56
GoogLeNet [2]	47.90	67.43	43.45	63.53	38.24	59.51	43.20	60.04
FACT [15]	**49.53**	67.96	44.63	64.19	39.91	60.49	44.69	64.21
DRDL [16]	48.91	66.71	**46.36**	64.38	**40.97**	60.02	**45.41**	63.70
NuFACT	48.90	**69.51**	43.64	**65.34**	38.63	**60.72**	43.72	**65.19**

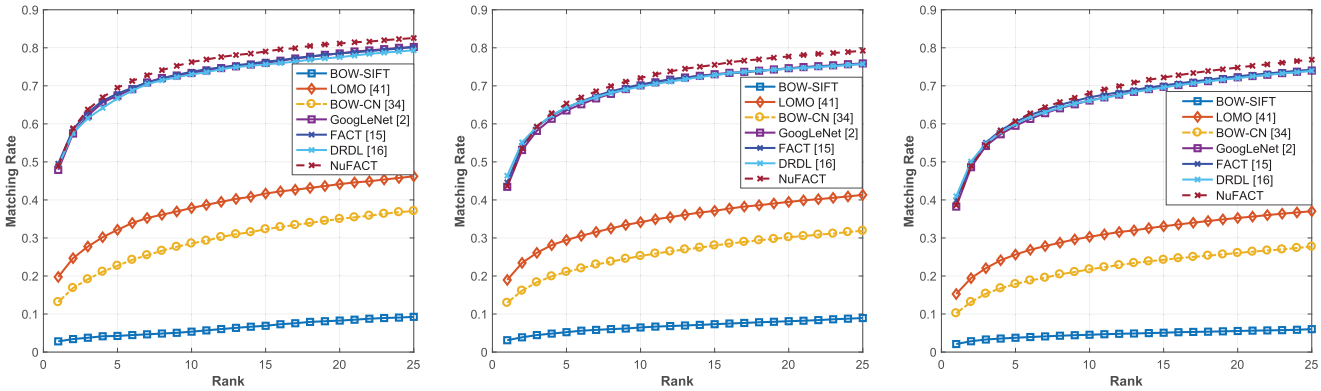

Fig. 10. The CMC curves of different methods on VehicleID. (a) Test size = 800. (b) Test size = 1600. (c) Test size = 2400.

fusion of color feature with semantic attributes can work better on VeRi. The other reason is that each vehicle in VeRi has many more images (64 images/vehicle) than the vehicles in VehicleID (8.4 images/vehicle). During training of the null space, more information such as different viewpoints, occlusions, and resolutions can be learned on VeRi. Thus, the NuFACT achieves greater improvement than FACT on the VeRi dataset.

D. Evaluation of Plate Verification

In this section, we compare the plate verification based on SNN with that based on the traditional texture features, i.e., SIFT [31]. The plate features obtained by the above two models are fused with the appearance features of the NuFACT model by the late fusion to evaluate the performance. The details of the two methods are as follows:

1) *NuFACT + Plate-SIFT*: This approach uses the hand-crafted SIFT as the basic representation. Then, the SIFT feature is quantized by the BOW model on the whole plate image. In the training phase, a codebook (size $k = 1000$) is learned on the training data of the VeRi dataset. During testing, the license plate image is extracted by the trained model as a 1000-D feature. Finally the plate feature and the appearance-based feature are integrated by late fusion.
2) *NuFACT + Plate-SNN*: This method adopts the SNN as the feature extractor for license plate images. During training, we first select over 100,000 plate pairs from the original 7,647 plates in the training set. Half of the pairs are from the same vehicles and are labeled with 1 as the positive samples; the other half are from different vehicles and are labeled with 0 as the negative samples. All samples are shuffled before training. The Caffe deep learning tool [15] is adopted to implement the SNN with the structure and parameters in Section V. The model is optimized by Stochastic Gradient Descent algorithm and converges after 60,000 iterations. Then, the output of the FC2 layer (1000-D) is extracted as the feature of the license plate images.

TABLE III
COMPARISON OF DIFFERENT MODELS FOR PLATE VERIFICATION

Methods	mAP	HIT@1	HIT@5
NuFACT + Plate-SIFT	42.48	75.27	90.41
NuFACT + Plate-SNN	**50.87**	**81.11**	**92.79**

Similar to appearance-based search, we estimate the similarity with Euclidean distance and perform the image-to-track search. The weights for late fusion are set to 0.86 and 0.14 for the NuFACT and the Plate-SNN models respectively.

Table III shows mAP, HIT@1, and HIT@5 on the VeRi dataset. The results show that the plate representation model learned by the deep neural network significantly outperformed the hand-crafted feature. Therefore, the features learned by SNN are more robust to uncertain environmental factors such as varied

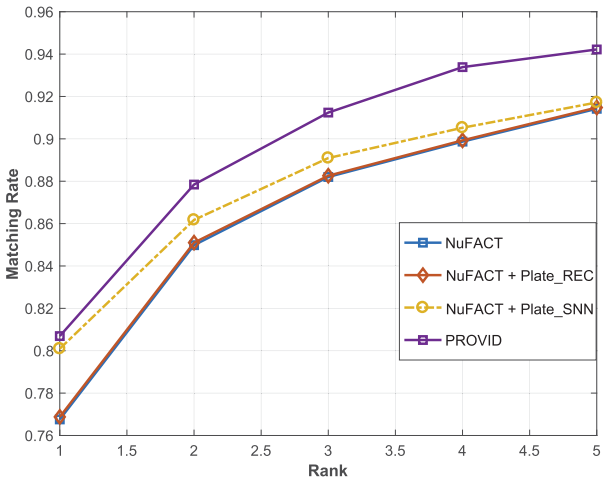

Fig. 11. The CMC curves of different methods.

lightning conditions and low resolution. This also demonstrates that the deep neural network has higher discriminative power especially with a large amount of training data. The effectiveness of the learned SNN is guaranteed by the use of sufficiently many license plate images.

E. Evaluation of Progressive Vehicle Re-Id

To evaluate the performance of the progressive search paradigm, we compare four methods on the VeRi dataset:

1) *NuFACT*: We utilize the NuFACT model to calculate the appearance similarities with the same settings as in Section VIII-C.
2) *NuFACT + Plate-SNN*: In this method, the NuFACT is first used to filter out the dissimilar vehicles by appearance. The late fusion scheme is then adopted to integrate the scores of the NuFACT model and Plate-SNN model for precise Re-Id. The weights for the NuFACT and the Plate-SNN models are set to 0.86 and 0.14 respectively as in Section VIII-D.
3) *NuFACT + Plate-REC*: This approach uses a commercial plate recognition tool (Plate-REC) to recognize the plate characters from the plate images for the accurate vehicle search. The weights for NuFACT and Plate-REC are set to 0.9 and 0.1 respectively as in Section VIII-D.
4) *PROVID*: This is the proposed progressive vehicle search framework, which fuses the scores of the NuFACT, Plate-SNN, and STR models. The Euclidean distance is adopted to compute the similarity between a query image and a test track. The NuFACT+Plate-SNN is obtained as introduced in Section VIII-D. The STR is computed with (6). Before late fusion, the similarity vectors are normalized to $(0, 1)$. Finally, the two vectors are added linearly to obtain the final scores. The weights are set to 0.85 and 0.15 for NuFACT+Plate-SNN and STR, respectively. Towards this end, the progressive vehicle Re-Id is achieved by comprehensively integrating the appearance features, license plate information, and spatiotemporal cues.

TABLE IV
COMPARISON OF DIFFERENT METHODS ON VERI DATASET

Methods	mAP	HIT@1	HIT@5
NuFACT	48.47	76.76	91.42
NuFACT + Plate-REC	48.55	76.88	91.48
NuFACT + Plate-SNN	50.87	81.11	92.79
PROVID	**53.42**	**81.56**	**95.11**

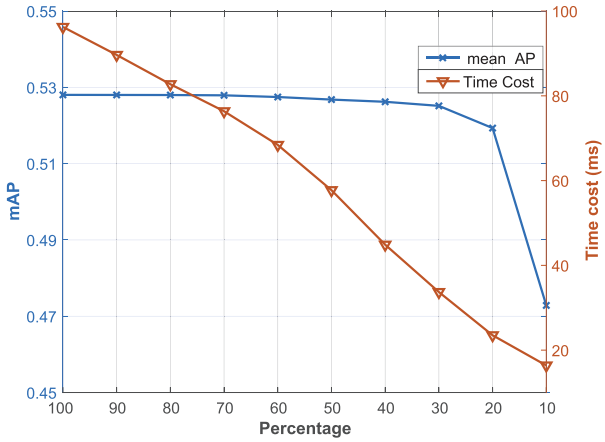

Fig. 12. The time cost and mAP under different top-K percent.

Fig. 11 shows the CMC curves of the progressive search. The mAP, HIT@1, and HIT@5 values are listed in Table IV. From the results, we can find that:

The results indicate that the proposed framework is effective for vehicle search with coarse filtering by appearance and precise matching by plate verification. The coarse filtering scheme can find most vehicles of similar shape, color, and type to the query vehicle, especially those with similar plate images. Moreover, after the filtering the vehicles, the framework can match the vehicles by license plate verification to eliminate the incorrect matches. The Plate-REC approach shows only negligible improvement because the recognition technique cannot achieve correct results under the unconstrained conditions. Furthermore, the PROVID framework outperforms all other tested approaches. In particular, the proposed framework can search the vehicles in the spatiotemporal space progressively in a close-to-far manner. The results validate the effectiveness of the PROVID framework as well as the significance of multi-modal data for vehicle search in large-scale urban surveillance.

In Fig. 13, we give some examples to compare efficacy of the proposed framework and our previous methods [22] on the VeRi dataset. For each query, the left three rows are the results of FACT, FACT+Plate-SNN, and FACT+Plate-SNN+STR in [22], the right three rows are the results of NuFACT, NuFACT+Plate-SNN, and the PROVID proposed in this paper. The three queries are hard cases in [22]. For example (a), the methods in [22] cannot return optimal results, even through the progressive search procedure, while the proposed PROVID can achieve excellent results in the top-five lists using only the appearance-based NuFACT model. This demonstrates the effectiveness and robustness of our NuFACT model in representing vehicle appearance.

Fig. 13. The top-5 search results on the VeRi dataset. For each query, the left three rows are the results of the FACT, FACT+Plate-SNN, and FACT+Plate-SNN+STR in [22], and the right three rows are the results of NuFACT, NuFACT+Plate-SNN, and PROVID proposed in this paper. The green box denotes a true positive, the red denotes a false positive. (Best seen in color.)

Example (b) shows the importance of the license plate verification in vehicle Re-Id. The vehicles with similar types and colors are found by the appearance features, but the correct results are not in the top results among the vehicles. Through the license plate verification, the target vehicles are matched precisely. From example (c), we can find that due to the low resolution and significant blur, the license plate verification may fail. Nevertheless, the target vehicles are found by the contextual information, i.e., the spatiotemporal similarity. These examples show the superior performance of the proposed PROVID framework compared to previous methods. However, the examples also reflect some limitations and difficulties of the system, which mainly come from three aspects: The first difficulty is caused by environmental factors. For example, varied illumination makes the same vehicle have very different colors especially in dark conditions. Moreover, the vehicle body under the sunlight can be very bright due to specular reflection. The second difficulty is caused by arbitrary camera settings. For example, the cameras in an urban surveillance system are not only installed in arbitrary locations, heights, and orientations but also with varied parameter settings, such as resolution, focal distance, and shutter speed. Therefore, the vehicle images captured by these cameras could contain significant blur, noise, and occlusion. The last difficulty is the ambiguity in the appearances of vehicles that are made by the same manufacturer and are of similar model and color. In this case, the license plate is the only information that can identify a vehicle. If the license plate is fake, occluded or removed, the proposed method might become invalid. However, even in these extreme conditions, PROVID can also provide valuable assistance in finding the target vehicle with the multi-modal information from urban surveillance.

In future work, we must explore a more discriminative and robust representation for vehicle appearance under unconstrained and uncertain surveillance environment, such dark illumination

or night scenes. In addition, license plate recognition techniques might be fused with the verification method by a multi-task learning framework, so the license plate information can be utilized more comprehensively for vehicle Re-Id.

F. Time Cost of the PROVID Framework

In our PROVID framework, we can select the top-K percent of outputs from the appearance-based filter and the license-plate-based search as the inputs of their subsequent procedures. To evaluate the mAP for different top-K percentages, we implement the PROVID framework on VeRi and reduce the percentage from 100% to 10%. To measure the time cost under the each percentage, we add 99,029 junk tracks to the original 2,021 test tracks to build a 50-time gallery. As shown in Fig. 12, we find that from the top-100% to top-30%, the mAP decreases marginally while the time cost decreases from 92.4 ms/query to 32.5 ms/query. PROVID can guarantee optimal accuracy and reduce the time cost by 64.8% by using the top 30% of outputs in each process as the inputs of the next step. This demonstrates that PROVID can significantly improve the precision and reduce the time cost of the instance-level vehicle search in large-scale urban surveillance.

IX. Conclusion

This paper proposes PROVID, a progressive vehicle re-identification framework based on deep learning which comprehensively exploits the multi-level features and multi-modal data in urban surveillance. We employ NFST to fuse the low-level features and the high-level attributes learned by deep CNN as the coarse filter. The precise search is achieved by license plate verification with Siamese neural network. Moreover, the proposed framework utilizes the spatiotemporal cues of the vehicles as the contextual information to re-rank the search results. Furthermore, we collect one of the most comprehensive dataset for vehicle Re-Id from practical traffic surveillance videos, which provides not only a sufficient number of vehicles but also license plates and contextual information. The extensive evaluations on the VeRi dataset show the excellent performance of the proposed PROVID framework.

References

[1] R. S. Feris *et al.*, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 28–42, Feb. 2012.

[2] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 3973–3981.

[3] B. C. Matei, H. S. Sawhney, and S. Samarasekera, "Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2011, pp. 3465–3472.

[4] N. Li, J. J. Jain, and C. Busso, "Modeling of driver behavior in real world scenarios using multiple noninvasive sensors," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1213–1225, Aug. 2013.

[5] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: A review," *Proc. Inst. Elect. Eng., Image Signal Process.*, vol. 152, no. 2, 2005, pp. 192–204.

[6] J. Zhang *et al.*, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transport. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[7] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 38, pp. 1–55, 2014.

[8] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[9] L. Xie, J. Wang, B. Zhang, and Q. Tian, "Fine-grained image search," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 636–647, May 2015.

[10] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.

[11] J. Meng, J. Yuan, J. Yang, G. Wang, and Y.-P. Tan, "Object instance search in videos via spatio-temporal trajectory discovery," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 116–127, Jan. 2016.

[12] H. D. Ma and W. Liu, "Progressive search paradigm for internet of things," *IEEE MultiMedia*, DOI: 10.1109/MMUL.2017.265091429.

[13] Y. Wen *et al.*, "An algorithm for license plate recognition applied to intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 830–845, Sep. 2011.

[14] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic license plate recognition (alpr): A state-of-the-art review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 311–325, Feb. 2013.

[15] X. C. Liu, W. Liu, H. D. Ma, and H. Y. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2016, pp. 1–6.

[16] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 2167–2175.

[17] D. Zapletal and A. Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops*, 2016, pp. 25–31.

[18] V. Kettnaker and R. Zabih, "Bayesian multi-camera surveillance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 1999, pp. 253–259.

[19] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Comput. Vis. Image Understand.*, vol. 109, no. 2, pp. 146–162, 2008.

[20] S. Sunderrajan and B. Manjunath, "Context-aware hypergraph modeling for re-identification and summarization," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 51–63, Jan. 2016.

[21] J. Xu, V. Jagadeesh, Z. Ni, S. Sunderrajan, and B. Manjunath, "Graph-based topic-focused retrieval in distributed camera network," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2046–2057, Dec. 2013.

[22] X. C. Liu, W. Liu, T. Mei, and H. D. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 869–884.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[25] A. Frome *et al.*, "Devise: A deep visual-semantic embedding model," in *Proc. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[26] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 3707–3715.

[27] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Surveys*, vol. 46, no. 38, 2014.

[28] J. Bromley *et al.* "Signature verification using a "siamese" time delay neural network," *Int. J. Pattern Recogn. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[29] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2005, pp. 539–546.

[30] C. Zhang, W. Liu, H. D. Ma, and H. Y. Fu, "Siamese neural network based gait recognition for human identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 2832–2836.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[32] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[33] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009.

[34] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.
[35] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 1–9.
[36] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[37] Y.-F. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue, "Null Foley–Sammon transform," *Pattern Recogn.*, vol. 39, no. 11, pp. 2248–2251, 2006.
[38] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2016, pp. 1239–1248.
[39] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors," *IEEE Trans. Comput.*, vol. C-100, no. 3, pp. 281–289, Mar. 1975.
[40] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[41] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015, pp. 2197–2206.
[42] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," CoRR, 2014, [Online]. Available: https://arxiv.org/abs/1405.3531, 2014.

Xinchen Liu (S'16) received the B.E. degree in computer science from Northwest A & F University, Shaanxi, China, in 2011. He is currently working toward the Ph.D. degree at the Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include multimedia content analysis and computer vision. Mr. Liu received the Best Student Paper Awards at ICME in 2016.

Wu Liu (S'14–M'15) received the B.E. degree from Shandong University, Shandong, China, in 2009, and the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently a Lecturer in Beijing Key Lab of Intelligent Telecommunications Software and Multimedia, Beijing University of Post and Telecommunication, Beijing. His research interests include multimedia information retrieval and computer vision. Dr. Liu received the Chinese Academy of Sciences Outstanding Ph.D. Thesis Award in 2016, and the Best Student Paper Awards at ICME in 2016.

Tao Mei (M'07–SM'11) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a Senior Researcher with Microsoft Research Asia. His current research interests include multimedia analysis and understanding, computer vision, and machine learning. He has authored or co-authored over 100 papers, 10 book chapters, and edited 4 books. He holds more than 40 U.S. and international patents (with 18 granted). Dr. Mei received a number of paper awards from prestigious multimedia journals and conferences, including the Best Paper Awards from *ACM Transactions on Multimedia Computing, Communications, and Applications* (2017), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2014), IEEE TRANSACTIONS ON MULTIMEDIA (2013), *ACM International Conference on Multimedia* (2007 and 2009), and so on. He is an Editorial Board Member of IEEE TRANSACTIONS ON MULTIMEDIA; *ACM Transactions on Multimedia Computing, Communications, and Applications*; *ACM Transactions on Intelligent Systems and Technology*; *Pattern Recognition*; and so on. He is the General Co-Chair of IEEE ICME 2019 and ICIMCS 2013, the Program Co-Chair of ACM Multimedia 2018, CBMI 2017, IEEE ICME 2015, IEEE MMSP 2015, and MMM 2013. He is a Fellow of IAPR and a Distinguished Scientist of ACM.

Huadong Ma (M'99–SM'16) received the B.S. degree in mathematics from Henan Normal University, Xinxiang, China, in 1984, the M.S. degree in computer science from Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang, China, in 1990, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1995. He is a Chang Jiang Scholar Professor, the Director of the Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, and the Executive Dean of the School of Computer Science with Beijing University of Posts and Telecommunications, Beijing. In 1998 and 1999, he was a Research Fellow with the United Nations University International Institute for Software Technology. From 1999 to 2000, he held a visiting position with the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI, USA. He was a Visiting Professor with The University of Texas at Arlington, Arlington, TX, USA, from July to September 2004 and with the Hong Kong University of Science and Technology from December 2006 to February 2007. He has published more than 200 papers and four books in his areas of interest. His current research focuses on multimedia system and networking, sensor networks, and Internet of Things. Prof. Ma is an Editorial Board Member of the IEEE TRANSACTIONS ON MULTIMEDIA and *Multimedia Tools and Applications*. He serves as the Chair of ACM SIGMOBILE CHINA. He received the National Funds for Distinguished Young Scientists in 2009.