Coarse-to-Fine: A RNN-Based Hierarchical Attention Model for Vehicle Re-identification

Xiu-Shen Wei[1(✉)], Chen-Lin Zhang[2], Lingqiao Liu[3], Chunhua Shen[3], and Jianxin Wu[2]

[1] Megvii Research Nanjing, Megvii Technology, Nanjing, China
weixs.gm@gmail.com
[2] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{zhangcl,wujx}@lamda.nju.edu.cn
[3] School of Computer Science, The University of Adelaide, Adelaide, Australia
{lingqiao.liu,chunhua.shen}@adelaide.edu.au

Abstract. Vehicle re-identification is an important problem and becomes desirable with the rapid expansion of applications in video surveillance and intelligent transportation. By recalling the identification process of human vision, we are aware that there exists a native hierarchical dependency when humans identify different vehicles. Specifically, humans always firstly determine one vehicle's coarse-grained category, *i.e.*, the car model/type. Then, under the branch of the predicted car model/type, they are going to identify specific vehicles by relying on subtle visual cues, *e.g.*, customized paintings and windshield stickers, at the fine-grained level. Inspired by the coarse-to-fine hierarchical process, we propose an end-to-end RNN-based Hierarchical Attention (RNN-HA) classification model for vehicle re-identification. RNN-HA consists of three mutually coupled modules: the first module generates image representations for vehicle images, the second hierarchical module models the aforementioned hierarchical dependent relationship, and the last attention module focuses on capturing the subtle visual information distinguishing specific vehicles from each other. By conducting comprehensive experiments on two vehicle re-identification benchmark datasets VeRi and VehicleID, we demonstrate that the proposed model achieves superior performance over state-of-the-art methods.

Keywords: Vehicle re-identification · Hierarchical dependency · Attention mechanism · Deep learning

The first two authors contributed equally to this work. This research was supported by NSFC (National Natural Science Foundation of China) under 61772256 and the program A for Outstanding Ph.D. candidate of Nanjing University (201702A010). Liu's participation was in part supported by ARC DECRA Fellowship (DE170101259).

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11362, pp. 575–591, 2019.
https://doi.org/10.1007/978-3-030-20890-5_37

1 Introduction

Vehicle re-identification is an important yet frontier problem, which aims at determining whether two images are taken from the same specific vehicle. It has diverse applications in video surveillance [30], intelligent transportation [37] and urban computing [40]. Moreover, vehicle re-identification has recently drawn increasing attentions in the computer vision community [19,20,23].

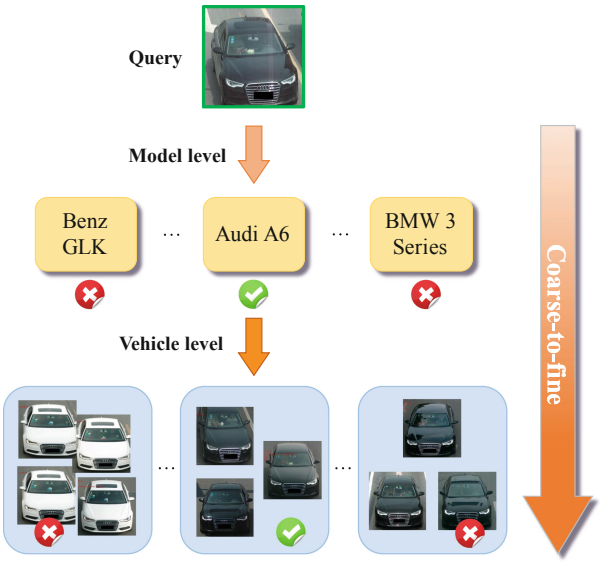

Fig. 1. Illustration of coarse-to-fine hierarchical information as a latent but crucial cue for vehicle re-identification. (Best viewed in color and zoomed in.) (Color figure online)

Compared with the classic person re-identification problem, vehicle re-identification could be more challenging as different specific vehicles can only be distinguished by slight and subtle differences, such as some customized paintings, windshield stickers, favorite decorations, etc. Nevertheless, there still conceals some latent but crucial information for handling this problem. As shown in Fig. 1, when humans identify different vehicles, they always follow a *coarse-to-fine* identification process. Specifically, we tend to firstly determine this specific vehicle belongs to which car model/type. The first step can eliminate many distractors, *i.e.*, vehicles with similar subtle visual appearances but belonging to the other different car models/types. In the following, within the candidate vehicle set of the same car model/type, humans will carefully distinguish different vehicles from each other by using these subtle visual cues. Apparently, there is a hierarchical dependency in this coarse-to-fine process, which is yet neglected by previous studies [16,19,20,23].

Motivated by such human's identification process, we propose a unified RNN-based Hierarchical Attention (RNN-HA) classification model for vehicle

re-identification. Specifically, as shown in Fig. 2, our RNN-HA consists of three main modules: *(1)* the representation learning module, *(2)* the RNN-based hierarchical module and *(3)* the attention module. The first module encodes the discriminative information of vehicle images into the corresponding deep convolutional descriptors. Then, the second RNN-based hierarchical module will mimic the coarse-to-fine identification process. Concretely, at the coarse-grained level classification (for car models), we first aggregate these deep descriptors by global average pooling (named as "image embedding vector $\boldsymbol{x}_1$" in Fig. 2), which is expected to retain global information contributing to the coarse-grained level. We then feed the image embedding vector $\boldsymbol{x}_1$ as the input at time step 1 into the RNN-based hierarchical module. The output at time step 1 is used for the coarse-grained level classification. More importantly, the same output is also treated as the source for generating an attention guidance signal which is crucial for the subsequent fine-grained level classification. At the fine-grained level classification (for specific vehicles), to capture subtle appearance cues, we leverage an attention module where the aforementioned attention guidance signal will evaluate which deep descriptors should be attended or overlooked. Based on the attention module, the original deep descriptors are transformed into the attended descriptors which reflect those image regions containing subtle discriminative information. After that, the attended descriptors are aggregated by global average pooling into the attention embedding vector $\boldsymbol{x}_2$ (cf. Fig. 2). Then, the attention embedding vector $\boldsymbol{x}_2$ is fed into RNN at time step 2 for the fine-grained level classification. In the evaluation, we employ the trained RNN-HA model as a feature extractor on test vehicle images (whose vehicle categories are disjoint with the training categories) to extract the outputs at time step 2 as the feature representations. For re-identification, these representations are first ℓ_2-normalized, and then the cosine acts as the similarity function for computing relative distances among different vehicles.

In experiments, we perform the proposed RNN-HA model on two vehicle re-identification benchmark datasets, *i.e.*, *VeRi* [19] and *VehicleID* [16]. Empirical results show that our RNN-HA model significantly outperforms state-of-the-art methods on both datasets. Furthermore, we also conduct ablation studies for separately investigating the effectiveness of the proposed RNN-based hierarchical and attention modules in RNN-HA.

In summary, our major contributions are three-fold:

- For vehicle re-identification, we propose a novel end-to-end trainable RNN-HA model consisting of three mutually coupled modules, especially two crucial modules (*i.e.*, the RNN-based hierarchical module and the attention module) which are tailored for this problem.
- Specifically, by leveraging powerful RNNs, the RNN-based hierarchical module models the coarse-to-fine category hierarchical dependency (*i.e.*, from car model to specific vehicle) beneath vehicle re-identification. Furthermore, the attention module is proposed for effectively capturing subtle visual appearance cues, which is crucial for distinguishing different specific vehicles (*e.g.*, two cars of the same car model).

– We conduct comprehensive experiments on two challenging vehicle re-identification datasets, and our proposed model achieves superior performance over competing previous studies on both datasets. Moreover, by comparing with our baseline methods, we validate the effectiveness of two proposed key modules.

2 Related Work

We briefly review two lines of related work: vehicle re-identification and developments of deep neural networks.

2.1 Vehicle Re-identification

Vehicle re-identification is an important application in video surveillance, intelligent transportation and public security [5,16,19,20,36]. It is a frontier research area in recent years with limited related studies in the literature.

Feris *et al.* [5] proposed an attribute-based approach for vehicle search in surveillance scenes. By classifying with different attributes, *e.g.*, colors and sizes, retrieval was performed on searching vehicles with such similar attributes in the database. In 2015, Yang *et al.* [36] proposed the large-scale car dataset *CompCar* for multiple car-related tasks, such as car model classification, car model verification and attribute prediction. However, the category granularity of the studying tasks in [25,36] just stayed in the car model level, which was not fine enough for the specific vehicle level. Very recently, Liu *et al.* [19] and Liu *et al.* [16] proposed their vehicle re-identification dataset *VeRi* and *VehicleID*, respectively. After that, vehicle re-identification started to gain attentions in the computer vision community.

Specifically, the *VeRi* dataset [19] contains over 50,000 images of 776 vehicles captured by twenty cameras in a road network. On this dataset, [19] proposed an appearance-based method named FACT combining both low-level features (*e.g.*, color and texture) and high-level semantic information extracted by deep neural networks. Later, [20] and [23] tried to deal with vehicle re-identification by using not only appearance information but also complex and hard-acquired spatio-temporal information. On the other hand, the *VehicleID* dataset [16] is a large-scale vehicle dataset containing about 26,000 different vehicles with 222,000 images. Liu *et al.* [16] proposed a mixed structured deep network with their coupled cluster loss to learn the relative distances of different vehicles on that dataset, which only depended on appearance cues.

In this paper, we attempt to use our RNN-HA model to deal with vehicle re-identification by depending on purely visual appearance cues. The reason is that appearance information is directly beneath vehicle images, and the images are easy to collect. Compared with that, complex and expensive spatio-temporal information is much harder to gather. For example, the two existing *large-scale* car/vehicle datasets, *e.g.*, *CompCar* [36] and *VehicleID* [16], only provide the appearance-based labels such as car types, models and identified annotations.

Whilst, they do not provide any spatio-temporal information. Moreover, our experimental results prove that it is practical to use only appearance information for accurate vehicle re-identification, cf. Table 2.

2.2 Deep Neural Networks

Deep convolutional neural networks (DCNNs) try to model the high-level abstractions of the visual data by using architectures composed of multiple non-linear transformations. Recent progresses in diverse computer vision applications, *e.g.*, large-scale image classification [14], object detection [2,18] and semantic segmentation [21], are made based on the developments of the powerful DCNNs.

On the other hand, the family of Recurrent Neural Networks (RNNs) including Gated Recurrent Neural Networks (GRUs) [3] and Long-Short Term Memory Networks (LSTMs) [8] have recently achieved great success in several tasks including image captioning [11,28,33,35], visual question answering [6,34], machine translation [26], etc. These works prove that RNNs are able to model the temporal dependency in sequential data and learn effective temporal feature representations, which inspires us to rely on RNNs for modeling the hierarchical coarse-to-fine characteristic (*i.e.*, from car model to specific vehicle) beneath the vehicle re-identification problem.

3 Model

We propose a novel RNN-based hierarchical attention (RNN-HA) classification model to solve the vehicle re-identification problem. In this section, we present our RNN-HA model by introducing its key constituent modules, *i.e.*, the representation learning module, the RNN-based hierarchical module and the attention module, respectively. The framework of our model is illustrated in Fig. 2.

3.1 Representation Learning Module

As shown in Fig. 2, the first module of RNN-HA is to learn a holistic image representation from an input vehicle image via traditional convolutional neural networks. The obtained deep image representations will be processed by the following modules. This representation learning module consists of multiple traditional convolutional layers (equipped with ReLU and pooling). However, different from previous vehicle re-identification methods (*e.g.*, [16,23]), we discard fully connected layers, and utilize convolutional activations for subsequent processing.

Concretely, the activations of a convolution layer can be formulated as an order-3 tensor T with $h \times w \times d$ elements, which includes a set of 2-D feature maps. These feature maps are embedded with rich spatial information, and are also known to obtain mid- and high-level information, *e.g.*, object parts [17,24]. From another point of view, these activations can alternatively be viewed as an

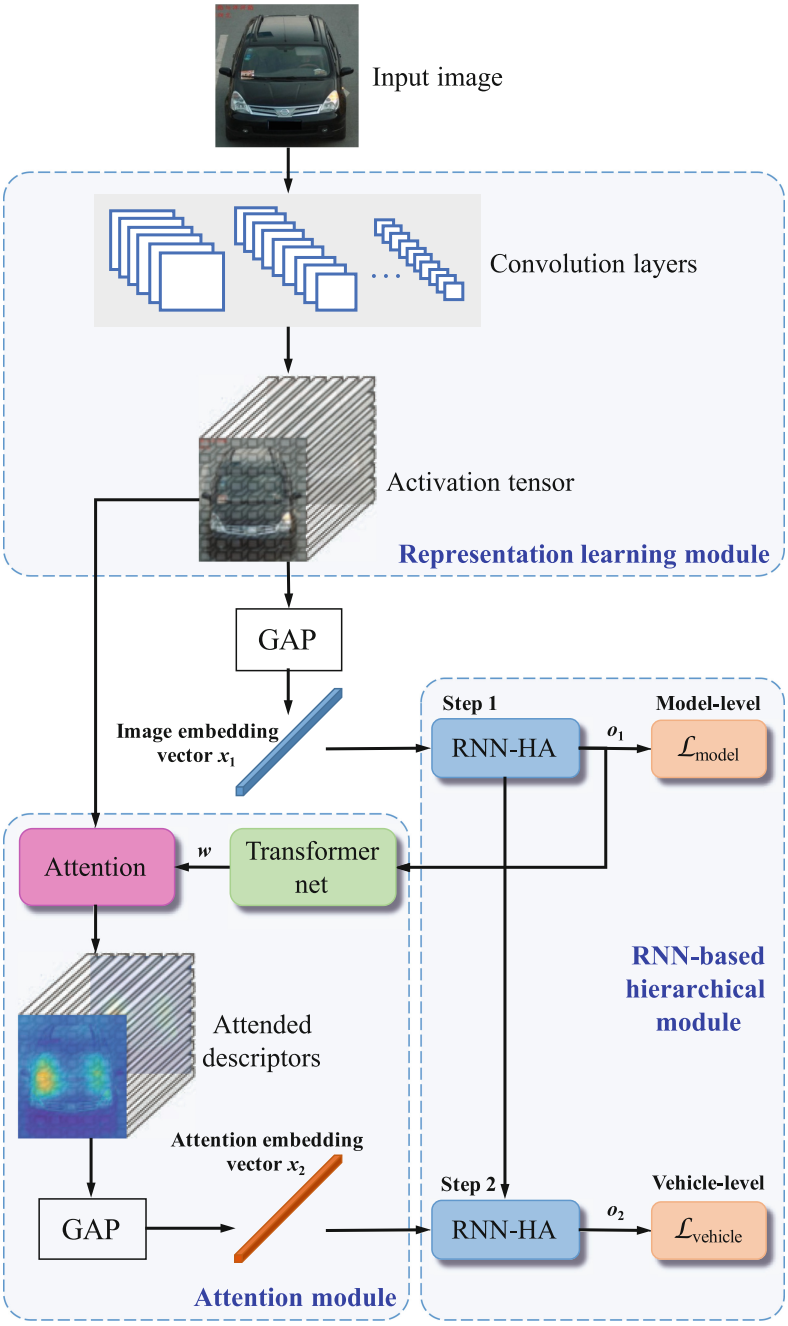

Fig. 2. Framework of the proposed RNN-HA model. Our model consists of three mutually coupled modules, *i.e.*, representation learning module, RNN-based hierarchical module and attention module. (Best viewed in color.) (Color figure online)

array of d-dimensional deep descriptors extracted at $h \times w$ locations. Apparently, these deep convolutional descriptors have more local/subtle visual information and more spatial information than those of the fully connected layer features.

After obtaining the activation tensor, at the coarse-grained classification level (*i.e.*, model-level), we directly apply global average pooling (GAP) upon these deep descriptors and regard the pooled representation as the "image embedding vector" $\boldsymbol{x}_1$. Then, $\boldsymbol{x}_1$ will be fed as the input at time step 1 into the RNN-based hierarchical module. Whilst, at the fine-grained classification level (*i.e.*, vehicle-level), these deep descriptors plus the intermediate output of the RNN-based hierarchical module will be used for generating the attended descriptors with the following attention embedding vector $\boldsymbol{x}_2$, cf. Fig. 2.

3.2 Crucial Modules in RNN-HA

In the following sections, we elaborate the RNN-based hierarchical module and the attention module which are two crucial modules of our RNN-HA model.

RNN-based Hierarchical Module. Recurrent Neural Network (RNN) [8] is a class of neural network that maintains internal hidden states to model the dynamic temporal behavior of sequences with arbitrary lengths through directed cyclic connections between its units. It can be considered as a hidden Markov model extension that employs non-linear transition function and is capable of modeling long/short term temporal dependencies.

We hereby employ RNN to capture the latent hierarchical label dependencies existing in vehicle re-identification. Concretely, we choose Gated Recurrent Units (GRUs) [3] as the gating mechanism to implement RNN. GRUs have fewer parameters than another popular gating mechanism, *i.e.*, LSTM [8], which makes GRU much easier to optimize. Moreover, evaluations by Chung *et al.* [4] found that when LSTM and GRU have the same amount of parameters, GRU slightly outperforms LSTM. Similar observations were also corroborated in [12].

GRU contains two gates: an update gate $\boldsymbol{z}$ and a reset gate $\boldsymbol{r}$. We follow the model used in [3]. Let σ be the sigmoid non-linear activation function. The GRU updates for time step t given inputs $\boldsymbol{x}_t$, $\boldsymbol{h}_{t-1}$ are:

$$\boldsymbol{z}_t = \sigma(W_{xz}\boldsymbol{x}_t + W_{hz}\boldsymbol{h}_{t-1} + b_z), \tag{1}$$

$$\boldsymbol{r}_t = \sigma(W_{xr}\boldsymbol{x}_t + W_{hr}\boldsymbol{h}_{t-1} + b_r), \tag{2}$$

$$\boldsymbol{n}_t = \tanh(W_{xg}\boldsymbol{x}_t + \boldsymbol{r}_t \odot W_{hg}\boldsymbol{h}_{t-1} + b_g), \tag{3}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{n}_t + \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1}. \tag{4}$$

Here, $\odot$ represents the product with a gate value, and various W matrices are learned parameters.

As shown in Fig. 2, we decompose the coarse-to-fine hierarchical classification problem into an ordered prediction path, *i.e.*, from car model to specific vehicle. The prediction path can reveal the hierarchical characteristic beneath this two-stage classification problem. Meanwhile, the probability of a path can be computed by the RNN model.

As aforementioned, since we aim to solve coarse-grained classification at the first step, global image features which represent global visual information are required. Thus, the image embedding vector $\boldsymbol{x}_1$ produced by simply global average pooling is used as the input at the first time step ($t = 1$) in the hierarchical module. Then, the output vector $\boldsymbol{o}_1$ at $t = 1$ is employed for computing the coarse-grained (model-level) classification loss, *i.e.*, $\mathcal{L}_{\text{model}}$. At the same time, $\boldsymbol{o}_1$ will be transformed via a transformer network (cf. the green sub-module in Fig. 2) into an attention guidance signal $\boldsymbol{w}$. The attention signal can guide the subsequent attention network to learn which deep descriptors should be attended when identifying different specific vehicles. The details of the attention network are presented in the next sub-section.

Now suppose we obtain a well-trained attention network, it could focus on these descriptors corresponding to subtle discriminative image regions (*e.g.*, customized paintings, favorite decorations, etc.), and neglect these descriptors corresponding to common patterns (*e.g.*, the similar headlights, car roofs, etc.). Based on the attentions on descriptors, we can obtain the attention embedding vector $\boldsymbol{x}_2$ which is also the input of the RNN-based hierarchical module at $t = 2$. The rest procedure at time step 2 is computing the loss in fine-grained classification (*i.e.*, vehicle-level classification) based on the output vector $\boldsymbol{o}_2$.

At last, the final loss of our RNN-HA is formed by the summation of both the coarse-grained (*i.e.*, model-level) and fine-grained (*i.e.*, vehicle-level) classification loss as:

$$\mathcal{L} = \mathcal{L}_{\text{model}} + \mathcal{L}_{\text{vehicle}} \,. \tag{5}$$

In our implementation, the traditional cross entropy loss function is employed in each loss branch. Note that, there is no tuning parameter in this equation, which reveals RNN-HA is generalized and not tricky.

Attention Module. After performing the coarse-grained classification, it is required to identify different specific vehicles at this fine-grained level. What distinguishes different vehicles from each other is the subtle visual information, such as customized paintings, windshield stickers, favorite decorations, etc. To capture these subtle but important cues, we propose to use an attention model to focus processing only on the attended deep descriptors reflecting those cues. Meanwhile, discarding those descriptors corresponding to common patterns, *e.g.*, the same car roofs of the same one car model, is also crucial for identifying specific vehicles.

To achieve these goals, we rely on the output $\boldsymbol{o}_1$ of the RNN-based hierarchical module at $t = 1$ by regarding it as the guidance signal for the subsequent attention learning. Specifically, we design a transformer network to transform $\boldsymbol{o}_1$ into a new space where it could play a role as the attention guidance signal $\boldsymbol{w}$. The transformer network contains two fully connected layers with a ReLU layer between them. After this transforming, $\boldsymbol{w}$ is utilized for evaluating which deep descriptors should be attended or overlooked.

Given an input image, $\boldsymbol{f}_{(i,j)} \in \mathbb{R}^d$ is the deep descriptor in the (i,j) spatial location in the activation tensor of the last convolutional layer, where $i \in \{1,2,\ldots,h\}$ and $j \in \{1,2,\ldots,w\}$. Based on the attention guidance signal $\boldsymbol{w}$, we can get the corresponding unnormalized attention scores $s_{(i,j)} \in \mathbb{R}^1$ of $\boldsymbol{f}_{(i,j)}$ by:

$$s_{(i,j)} = g(\boldsymbol{w}^\top \boldsymbol{f}_{(i,j)}), \tag{6}$$

where $g(\cdot)$ is the softplus function, *i.e.*, $g(x) = \ln(1+\exp(x))$. Since we only aim to concern with the relative importance of the deep descriptors within an image, we further normalize the attention scores into the $[0,1]$ interval for aggregating the descriptors:

$$a_{(i,j)} = \frac{s_{(i,j)} + \epsilon}{\sum_i \sum_j (s_{(i,j)} + \epsilon)}, \tag{7}$$

where $a_{(i,j)}$ is the normalized attention score, and ϵ is a small constant set to 0.1 in our experiments.

Thus, different deep descriptors will receive the corresponding (normalized) attention scores depending on how much attention paid to itself. Finally, we can get the attended feature representation $\hat{\boldsymbol{f}}_{(i,j)} \in \mathbb{R}^d$ by applying $a_{(i,j)}$ to $\boldsymbol{f}_{(i,j)}$ as follows:

$$\hat{\boldsymbol{f}}_{(i,j)} = a_{(i,j)} \odot \boldsymbol{f}_{(i,j)}, \tag{8}$$

where $\odot$ is the element-wise multiplication. Then, after global average pooling employed in $\hat{\boldsymbol{f}}_{(i,j)}$, we can obtain the attention embedding vector:

$$\boldsymbol{x}_2 = \frac{1}{h \times w} \sum_i \sum_j \hat{\boldsymbol{f}}_{(i,j)}. \tag{9}$$

Then, $\boldsymbol{x}_2 \in \mathbb{R}^d$ will be fed as the input into the RNN-based hierarchical module at $t = 2$. At last, by treating different vehicles as classification categories, the output $\boldsymbol{o}_2$ at $t = 2$ is used for recognizing vehicles at the fine-grained level. As each module is differentiable, the whole RNN-HA model is end-to-end trainable.

Additionally, in theory, the proposed RNN-based hierarchical module can be composed of more time steps, *i.e.*, $t > 2$. Since the vehicle datasets used in experiments have two-level hierarchical labels, we use our RNN-HA with $t = 2$ for the vehicle re-identification problem. Beyond that, the proposed RNN-HA model can also deal with the hierarchical fine-grained recognition problem, *e.g.*, [9], which reveals the generalization usage of our model.

4 Experiments

In this section, we first describe the datasets and evaluation metric used in experiments, and present the implementation details. Then, we report vehicle re-identification results on two challenging benchmark datasets.

4.1 Datasets and Evaluation Metric

To evaluate the effectiveness of our proposed RNN-HA model, we conduct experiments on two benchmark vehicle re-identification datasets, *i.e.*, *VeRi* [19] and *VehicleID* [16].

The *VeRi* dataset contains more than 50,000 images of 776 vehicles with identity annotations and car types. The car types in *VeRi* include the following ten kinds: "sedan", "SUV", "van", "hatchback", "MPV", "pickup", "bus", "truck" and "estate". The *VeRi* dataset is split into a training set consisting of 37,778 images of 576 vehicles and a test set of 11,579 images belonging to 200 vehicles. In the evaluation, a subset of 1,678 images in the test set are used as the query images. The test protocol in [19,20] recommends that evaluation should be conducted in an image-to-track fashion, in which the image is used as the query, while the gallery consists of tracks of the same vehicle captured by other cameras. The mean average precision (mAP), top-1 and top-5 accuracy (CMC) are chosen as the evaluation metric for the *VeRi* dataset.

Compared with *VeRi*, *VehicleID* is a large-scale vehicle re-identification dataset, which has a total of 26,267 vehicles with 221,763 images, and 10,319 vehicles are labeled with models such as "Audi A6", "Ford Focus", "Benz GLK" and so on. There are 228 car models in total. *VehicleID* 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. For *VehicleID*, the image-to-image search is conducted because each vehicle is captured in one image by one camera. For each test dataset (size = 800, 1,600 and 2,400), one image of each vehicle is randomly selected into the gallery set, and all the other images are query images. We repeat the above processing for ten times, and report the average results. Following the previous work [16], the evaluation metric for *VehicleID* is top-1, top-5 accuracy (CMC).

4.2 Implementation Details

In our experiments, vehicles' identity annotations of both datasets are treated as the vehicle-level classification categories, while the car types of *VeRi* and the car models of *VehicleID* are regarded as the model-level classification categories for these two datasets, respectively. Note that, for both datasets, the vehicle-level classification categories of training and testing are disjoint. All images are of 224×224 image resolution. After training RNN-HA, we employ our model as a feature extractor for extracting test image features. Specifically, the output $\boldsymbol{o}_2$ at time step 2 of our RNN-HA is the acquired image representation. In the evaluation, $\boldsymbol{o}_2$ of test images are firstly ℓ_2-normalized, and then the cosine distance acts as the similarity function in re-identification.

For fair comparisons, following [16], we adopt *VGG_CNN_M_1024* [1] as the base model of our representation learning module. For the RNN-based module, the number of hidden units in GRU is 1,024, and the zero vector is the hidden state input at time step 0. During training, we train our unified RNN-HA in an end-to-end manner by employing the RMSprop [29] optimization method with

its default parameter settings to update model parameters. The learning rate is set to 0.001 at the beginning, and five epochs later, it is reset to 0.0001. The batch size is 64. We implement our model by the open-source library PyTorch.

4.3 Main Results

We present the main vehicle re-identification results by firstly introducing state-of-the-arts and our baseline methods, then following the comparison results on two datasets.[1]

Comparison Methods. We compare nine vehicle re-identification approaches which are evaluated on *VehicleID* and *VeRi*. The details of these approaches are introduced as follows. Among them, many state-of-the-art methods, *e.g.*, [13,16,42], also employ both vehicle-level and model-level supervision, which has the same formation as ours.

- *Local Maximal Occurrence Representation (LOMO)* [15] is the state-of-the-art text features for person re-identification. We follow the optimal parameter settings given in [15] when adopting LOMO on these two vehicle re-identification datasets.
- *Color based feature (BOW-CN)* [38] is a benchmark method in person re-identification, which applies the Bag-of-Words (BOW) model [10,22] with Color Name (CN) features [31]. By following [38], a 5,600-d BOW-CN feature is obtained as the color based feature.
- *Semantic feature learned by CNN (GoogLeNet)* adopts the GoogLeNet model [27] which is fine-tuned on the CompCars dataset [36] as a powerful feature extractor for high-level semantic attributes of the vehicle appearance. The image feature of this method is 1,024-d extracted from the last pooling layer of GoogLeNet.
- *Fusion of Attributes and Color features (FACT)* [19] is proposed for vehicle re-identification by combining deeply learned visual feature from GoogLeNet [27], BOW-CN and BOW-SIFT feature to measure only the visual similarity between pairs of query images.
- *Siamese-Visual* [23] relys on a siamese-based neural network which has a symmetric structure to learn the similarity between a query image pair with appearance information.
- *Triplet Loss* [32] is adopted in vehicle re-identification by learning a harmonic embedding of each input image in the Euclidean space that tends to maximize the relative distance between the matched pair and the mismatched pair.
- *Coupled Cluster Loss (CCL)* [16] is proposed for improving triplet loss by replacing the single positive/negative input sample by positive/negative images sets, which can make the training phase more stable and accelerate the convergence speed.

[1] Note that, in Tables 1 and 2, we only compare with the appearance-based methods for fair comparisons.

Table 1. Comparison of different methods on *VeRi* [19].

Methods	mAP	Top-1	Top-5
LOMO [15]	9.64	25.33	46.48
BOW-CN [38]	12.20	33.91	53.69
GoogLeNet [36]	17.89	52.32	72.17
FACT [19]	18.75	52.21	72.88
Siamese-Visual [23]	29.48	41.12	60.31
VAMI [42]	50.13	77.03	90.82
FC-HA (w/o RNN)	47.19	61.56	76.88
RNN-H w/o attention	48.92	63.28	78.82
Our RNN-HA	52.88	66.03	80.51
Our RNN-HA (ResNet)	**56.80**	**74.79**	**87.31**

- *Mixed Diff+CCL* [16] adopts a mixed network structure with a coupled cluster loss to learn the relative distances of different vehicles. Note that, the most different point between it and our model is that it used these two level categories in a simple parallel fashion, while we formulate them into a hierarchical fashion. Experimental results could validate the effectiveness of our proposed hierarchical model, and justify the hierarchical fashion should be the optimal option.
- *CLVR* [13] consists of two branches in the cross-modality paradigm, where one branch is for the vehicle model level and the other is for the vehicle ID level. Again, the two level categories in their model are not formulated into a hierarchical fashion like ours.
- *VAMI* [42] proposes a viewpoint-aware attentive multi-view inference model by leveraging the cues of multiple views to deal with vehicle re-identification.

Additionally, to investigate the impacts of the various modules in our end-to-end framework, we analyze the effects of the RNN-based hierarchical module and the attention module by conducting experiments on two baselines:

- *FC-HA (w/o RNN)* replaces the RNN hierarchical module by simply employing traditional fully-connected layers as direct transformation, but keeps the attention mechanism.
- *RNN-H w/o attention* keeps the hierarchical module, but removes the attention module from our proposed RNN-HA model. Specifically, the inputs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ at $t = 1$ and $t = 2$ are both the representations global average pooled by these deep convolutional descriptors.

Comparison Results on *VeRi*. Table 1 presents the comparison results on the *VeRi* dataset. Our proposed RNN-HA model achieves 52.88% mAP, 77.03% top-1 accuracy and 90.91% top-5 accuracy on *VeRi*, which significantly outperforms the other state-of-the-art methods. These results validate the effectiveness of

Table 2. Comparison of different methods on the *VehicleID* dataset [16].

Methods	Test size = 800		Test size = 1,600		Test size = 2,400	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
LOMO [15]	19.7	32.1	18.9	29.5	15.3	25.6
BOW-CN [38]	13.1	22.7	12.9	21.1	10.2	17.9
GoogLeNet [36]	47.9	67.4	43.5	63.5	38.2	59.5
FACT [19]	49.5	67.9	44.6	64.2	39.9	60.5
Triplet Loss [32]	40.4	61.7	35.4	54.6	31.9	50.3
CCL [16]	43.6	64.2	37.0	57.1	32.9	53.3
Mixed Diff+CCL [16]	49.0	73.5	42.8	66.8	38.2	61.6
CLVR [13]	62.0	76.0	56.1	71.8	50.6	68.0
VAMI [42]	63.1	83.3	52.8	75.1	47.3	70.3
FC-HA (w/o RNN)	56.7	74.5	53.6	70.6	48.6	66.3
RNN-H w/o attention	64.5	78.8	62.4	75.9	59.0	74.2
Our RNN-HA	68.8	81.9	66.2	79.6	62.6	77.0
Our RNN-HA (672)	74.9	85.3	71.1	82.3	68.0	81.4
Our RNN-HA (ResNet+672)	**83.8**	**88.1**	**81.9**	**87.0**	**81.1**	**87.4**

the proposed model. Moreover, RNN-HA has a gain of 5.69% mAP and 15.47% top-1 accuracy comparing with the FC-HA baseline method, which proves the effectiveness of the RNN-based hierarchical design. Also, compared with "RNN-H w/o attention", RNN-HA achieves a gain of 3.96% mAP and 11.75% top-1 accuracy. It justifies our proposed attention module when identifying different specific vehicles at the fine-grained classification level.

In addition, to further improve the re-identification accuracy, we simply replace the *VGG_CNN_M_1024* model of the representation learning module with *ResNet-50* [7]. Our modified model is denoted as "RNN-HA (ResNet)", which obtains 56.80% mAP, 80.79% top-1 accuracy and 92.31% top-5 accuracy on *VeRi*.

Comparison Results on *VehicleID*. For the large-scale dataset, *VehicleID*, we report the comparison results in Table 2. On different test settings (*i.e.*, test size = 800, 1,600 and 2,400), our proposed RNN-HA achieves the best re-identification performance on this large-scale dataset. An interesting observation in both tables is that the FC-HA baseline method outperforms all the Siamese or triplet training methods on both *VeRi* and *VehicleID*. It is consistent with the observations in recently most successful person re-identification approaches, *e.g.*, [39,41]. These approaches argue that a classification loss is superior for the re-identification task, while the triplet loss or siamese-based nets perform unsatisfactorily due to its tricky training example sampling strategy.

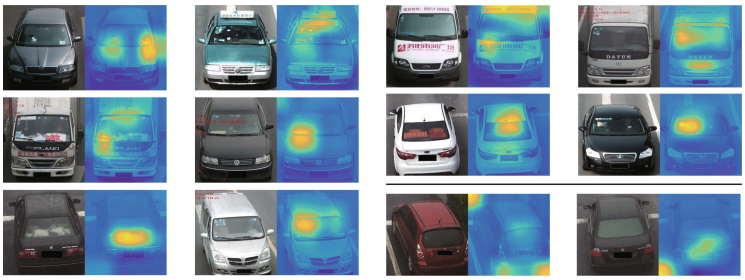

Fig. 3. Examples of the attention maps on *VehicleID*. The brighter the region, the higher the attention scores. (Best viewed in color and zoomed in.)

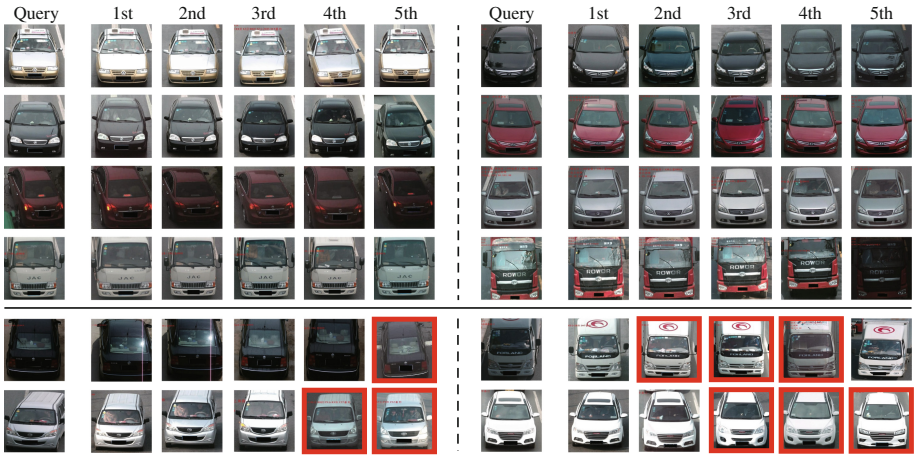

Fig. 4. The top-5 re-identification results on the large-scale *VehicleID* dataset. The upper eight examples are successful, while the lower four examples are failure cases. Red boxes denote false positives. (Best viewed in color and zoomed in.) (Color figure online)

From the qualitative perspective, Fig. 3 shows the learned attention maps (*i.e.*, $a_{(i,j)}$ in Eq. 7) of several random sampled test vehicle images. We can find that the attended regions accurately correspond to these subtle and discriminative image regions, such as windshield stickers, stuffs placed behind windshield or rear windshield, and customized paintings. In addition, Fig. 4 presents several re-identification results returned by our RNN-HA on *VehicleID*.

Furthermore, on this large-scale dataset, we also use input images with a high image resolution, *i.e.*, 672×672, since higher resolution could benefit to learn a more accurate attention map. The RNN-HA model with the 672×672 resolution is denoted by "RNN-HA (672)". Apparently, it improves the re-identification

top-1 accuracy by 5–6%. Besides, based on the high resolution, we also modify RNN-HA by equipping it with *ResNet-50*, and report its results as "RNN-HA (ResNet+672)" in Table 2. On such challenging large-scale dataset, even though we only depend on appearance information, our model could achieve 83.8% top-1 accuracy, which reveals its effectiveness in real-life applications.

5 Conclusion

In this paper, we noticed there is a coarse-to-fine hierarchical dependency natively beneath the vehicle re-identification problem, *i.e.*, from coarse-grained (car models/types) to fine-grained (specific vehicles). To model this important hierarchical dependent relationship, we proposed a unified RNN-based hierarchical attention (RNN-HA) model consisting of three mutually coupled modules. In RNN-HA, after obtaining the deep convolutional descriptors generated by the first representation learning module, we leveraged the powerful RNNs and further designed a RNN-based hierarchical module to mimic such hierarchical classification process. Moreover, at the fine-grained level, the attention module was developed to capture subtle appearance cues for effectively distinguish different specific vehicles.

On two benchmark datasets, *VeRi* and *VehicleID*, the proposed RNN-HA model achieved the best re-identification performance. Besides, extensive ablation studies demonstrated the effectiveness of individual modules of RNN-HA. In the future, it is promising to further improve the re-identification performance by incorporating more attributes information into our proposed RNN-HA framework.

References

1. Chatfield, K., Simonyan, K., Vedaldi, A.: Return of the devil in the details: delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
2. Chen, C., Liu, M.-Y., Tuzel, O., Xiao, J.: R-CNN for small object detection. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10115, pp. 214–230. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54193-8_14
3. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP, pp. 1724–1735, October 2014
4. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
5. Feris, R.S., Siddiquie, B., Zhai, Y., Datta, A., Brown, L.M., Pankanti, S.: Large-scale vehicle detection, indexing, and search in urban surveillance videos. IEEE TMM **14**(1), 28–42 (2015)
6. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? Dataset and methods for multilingual image question answering. In: NIPS, pp. 2296–2304, December 2015
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778, June 2016
8. Hochreiter, S., Schmidhuber, J.: Long shot-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

9. Hou, S., Feng, Y., Wang, Z.: VegFru: a domain-specific dataset for fine-grained visual categorization. In: ICCV, pp. 541–549, October 2017
10. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR, pp. 3304–3311, June 2010
11. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: fully convolutional localization networks for dense captioning. In: CVPR, pp. 4565–4574, June 2016
12. Jozefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: ICML, pp. 2342–2350, July 2015
13. Kanac, A., Zhu, X., Gong, S.: Vehicle re-identification by fine-grained cross-level deep learning. In: BMVC, pp. 770–781, September 2017
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105, December 2012
15. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR, pp. 2197–2206, June 2015
16. Liu, H., Tian, Y., Wang, Y., Pang, L., Huang, T.: Deep relative distance learning: tell the difference between similar vehicles. In: CVPR, pp. 2167–2175, June 2016
17. Liu, L., Shen, C., van den Hengel, A.: Cross-convolutional-layer pooling for image recognition. IEEE TPAMI **39**(11), 2305–2313 (2016)
18. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
19. Liu, X., Lin, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban suveillance videos. In: ICME, pp. 1–6, July 2016
20. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 869–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_53
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440, June 2015
22. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. IJCV **105**(3), 222–245 (2013)
23. Shen, Y., Xiao, T., Li, H., Yi, S., Wang, X.: Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals. In: ICCV, pp. 1900–1909, October 2017
24. Singh, B., Han, X., Wu, Z., Davis, L.S.: PSPGC: part-based seeds for parametric graph-cuts. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 360–375. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16811-1_24
25. Sochor, J., Herout, A., Havel, J.: BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition. In: CVPR, pp. 3006–3015, June 2016
26. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS, pp. 3104–3112, December 2014
27. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR, pp. 1–9, June 2015
28. Tan, Y.H., Chan, C.S.: phi-LSTM: a phrase-based hierarchical LSTM model for image captioning. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10115, pp. 101–117. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54193-8_7
29. Tieleman, T., Hinton, G.E.: Neural networks for machine learning. Coursera (Lecture 65 - RMSprop)
30. Varela, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. IEEE Trans. Vis. Image Signal Process. **152**(2), 192–204 (2005)

31. Weijier, J.V.D., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE TIP **18**(7), 1512–1523 (2009)
32. Weijier, J.V.D., Schmid, C., Verbeek, J., Larlus, D.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recogn. **48**(10), 2993–3003 (2015)
33. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: What value to explicit high level concepts have in vision to language problems? In: CVPR, pp. 203–212, June 2016
34. Wu, Q., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Ask me anything: free-form visual question answering based on knowledge from external sources. In: CVPR, pp. 4622–4630, June 2016
35. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057, July 2015
36. Yang, L., Luo, P., Loy, C.C., Tang, X.: A large-scale car dataset fro fine-grained categorization and verification. In: CVPR, pp. 3973–3981, June 2015
37. Zhang, J., Wang, F.Y., Lin, W.H., Xu, X., Chen, C.: Data-driven intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. **12**(4), 1624–1639 (2011)
38. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV, pp. 1116–1124, December 2015
39. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
40. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**(38), 1–55 (2014)
41. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person re-identification. arXiv preprint arXiv:1611.05666 (2016)
42. Zhou, Y., Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: CVPR, pp. 6489–6498, June 2018