# 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], Jianxin Wu[2]**

1 Megvii Research Nanjing, Megvii Technology Ltd. (Face++), China

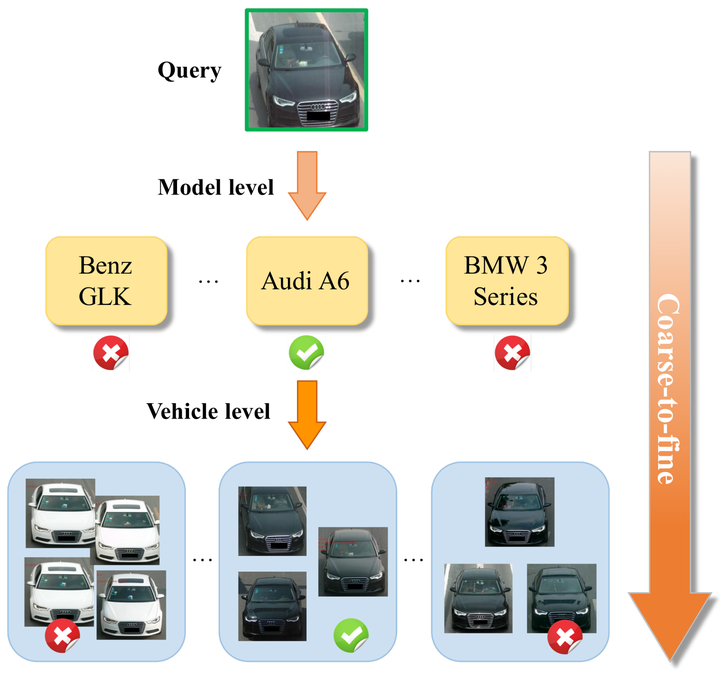
2 National Key Laboratory for Novel Software Technology, Nanjing University, China

3 The University of Adelaide, Adelaide, Australia



ACCV2018 2 – 6 December 2018 Perth Western Australia


## 1 – Motivation

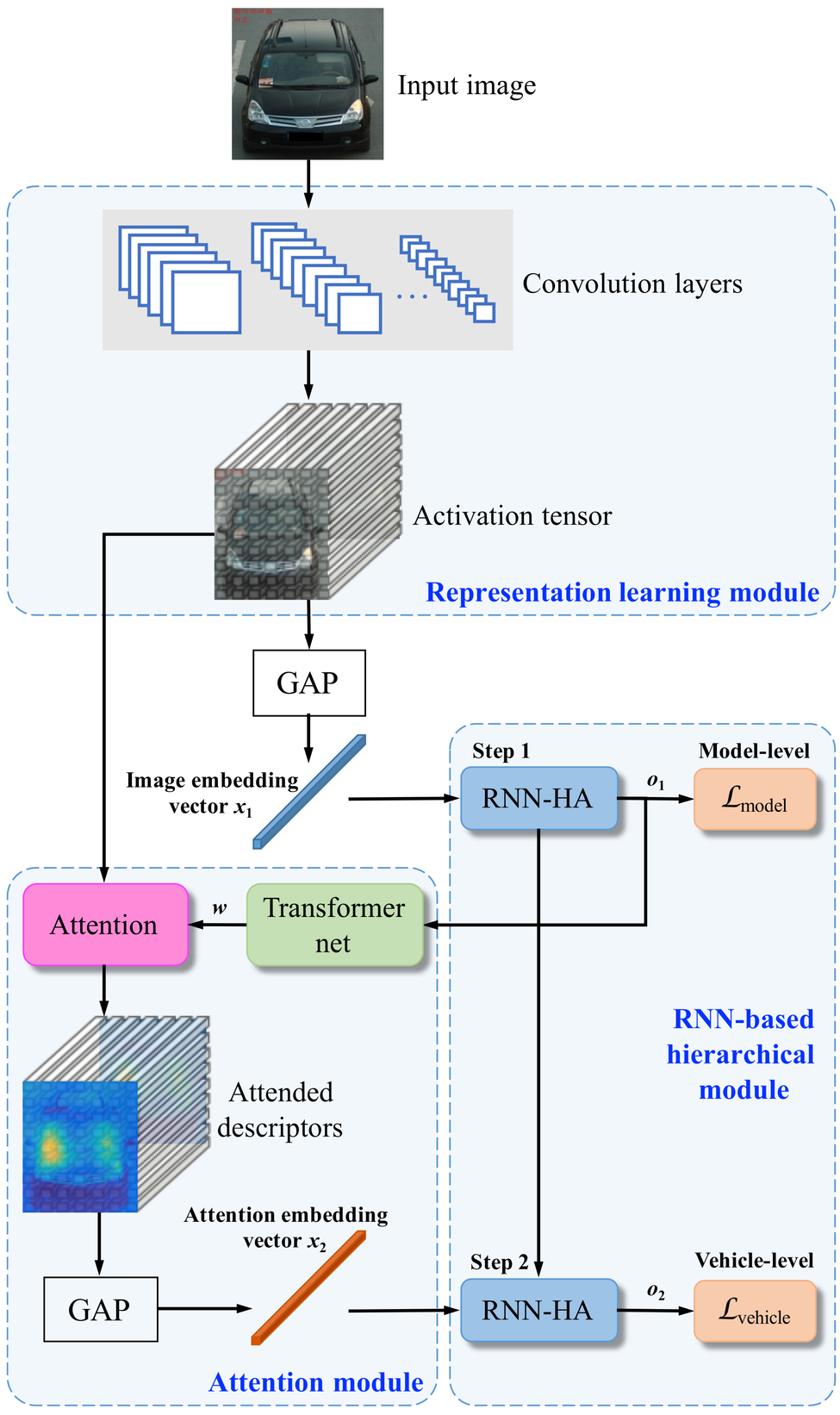



Humans always firstly determine one vehicle's coarse-grained category, *i.e.*, the car model/type. Then, under the branch of the predicted car model/type, they are going to identify specific vehicles by relying on subtle visual cues, *e.g.*, customized paintings and windshield stickers, at the fine-grained level.

## 3 – Contributions

- ✓ We propose a novel end-to-end trainable RNN-HA model consisting of three mutually coupled modules, especially the RNN-based hierarchical and attention modules which are tailored for this problem.
- ✓ Specifically, 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.
- ✓ 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 – The proposed method: RNN-HA




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.

## 4 – Experiments

Two benchmark vehicle re-identification datasets: *Veri* and *VehicleID*

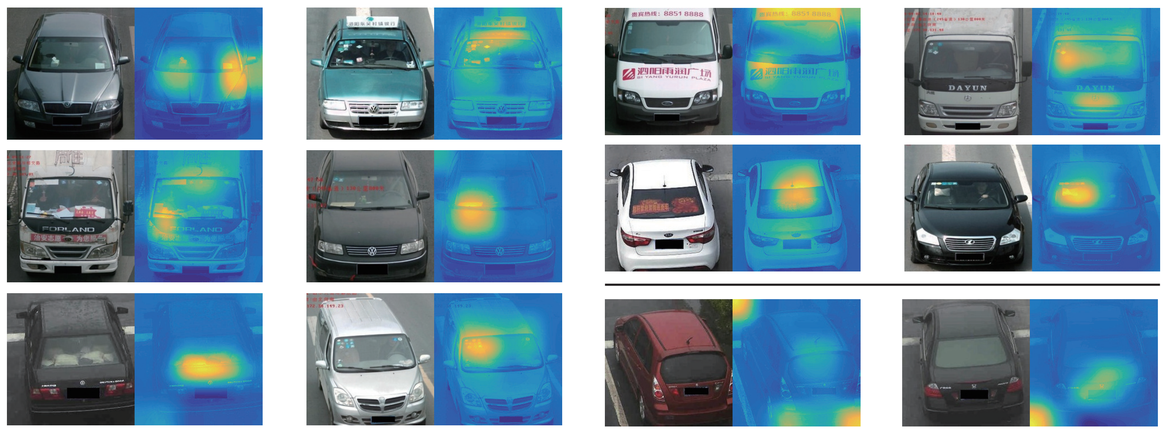
Quantitative results on the *Veri* dataset [Ref 1]:

| 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** |

Quantitative results on the *VehicleID* dataset [Ref2]:

| 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** |

Qualitative results of visualization of the attention maps on *VehicleID*



## 5 – References


[Ref 1] Liu, X. et al., *Large-scale vehicle re-identification in urban surveillance videos*. ICME'16.

[Ref 2] Liu, H. et al., *Deep relative distance learning: Tell the difference between similar vehicles*. CVPR'16.