A Dual-Path Model With Adaptive Attention For Vehicle Re-Identification

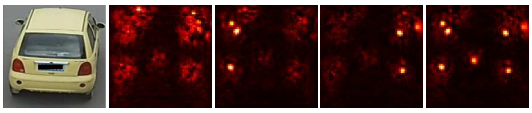
Pirazh Khorramshahi[1], Amit Kumar[1], Neehar Peri[1], Sai Saketh Rambhatla[1], Jun-Cheng Chen[2] and Rama Chellappa[1]
[1]Center for Automation Research, UMIACS, University of Maryland, College Park
[2]Research Center for Information Technology Innovation, Academia Sinica
{pirazhkh, akumar14, peri, rssaketh, rama}@umiacs.umd.edu, pullpull@citi.sinica.edu.tw

Abstract

In recent years, attention models have been extensively used for person and vehicle re-identification. Most re-identification methods are designed to focus attention on key-point locations. However, depending on the orientation, the contribution of each key-point varies. In this paper, we present a novel dual-path adaptive attention model for vehicle re-identification (AAVER). The global appearance path captures macroscopic vehicle features while the orientation conditioned part appearance path learns to capture localized discriminative features by focusing attention on the most informative key-points. Through extensive experimentation, we show that the proposed AAVER method is able to accurately re-identify vehicles in unconstrained scenarios, yielding state of the art results on the challenging dataset VeRi-776. As a byproduct, the proposed system is also able to accurately predict vehicle key-points and shows an improvement of more than 7% over state of the art. The code for key-point estimation model is available at https://github.com/Pirazh/Vehicle_Key_Point_Orientation_Estimation

1. Introduction

Vehicle re-identification (re-id) refers to the task of retrieving all images of a particular vehicle identity in a large gallery set, composed of vehicle images taken from varying orientations, cameras, time and locations. Accurately re-identifying vehicles from images and videos, is of great interest in surveillance and intelligence applications. In contrast to vehicle recognition which aims to identify the make and model of the vehicle, vehicle re-id is concerned with identifying specific vehicle instances. This task is extremely challenging as vehicles with different identities can be of the same make, model and color, and thus it is challenging for a Deep Convolutional Neural Network (DCNN) to make accurate predictions. In this paper, we present a novel algorithm driven by adaptive attention for re-identifying vehicles from still images without using information from other sources such as time and location.



(a) Front (b) Left (c) Right (d) Rear

Figure 1: Heatmaps grouped as suggested in [23]. Attention to all subgroup of key-points leads to erroneous results. Although, only the rear of the car is visible, contributions from frontal key-points are non-zero.

The similar task of person re-id aims at re-identifying humans appearing in different cameras. While visual appearance models work reasonably well for person re-id, the same techniques fail to differentiate vehicles due to the lack of highly discriminating features. Person re-id models are not heavily reliant on facial features as they also learn discriminating features based on clothing and accessories. However, vehicle re-id poses a new set of challenges. Different vehicle identities can have similar colors and shapes especially those coming from the same manufacturer with a particular model, trim and year. Subtle cues such as different wheel patterns and custom logos might be unavailable in the global appearance features. Therefore, it is important that vehicle re-id model learns to focus on different parts of the vehicles while making a decision. Previous works in person re-id such as [25] have used attention models with human key-points as regions of attention and have shown significant improvement in performance. Similarly, methods such as [23] have used vehicle key-points to learn attention maps for each of the 20 key-points defined by [23]. The system proposed by Wang *et al.* [23] grouped key-points into four groups corresponding to front, rear, left and right.

However, not all key-points provide discriminating information and their respective contributions depend on the orientation of the vehicle. For instance, in Figure 1a we observe that the key-points from the front of the car incor-

rectly influence the attention of the model as the front of the car is not visible. Hence, paying attention to all the key-points, as suggested in [23], can lead to erroneous results. The proposed method tackles the problem of false attention by adaptively choosing the key-points to focus on, based on the orientation of the vehicle hence, providing complementary information to global appearance features. In this work, terms with same connotation, *path*, *stream* and *branch*, have been used interchangeably.

In the proposed method, the first stream is a DCNN trained to extract discriminative global appearance features for each vehicle identity. However, this stream often fails to extract subtle features necessary to distinguish similar vehicles. Therefore, a second path composed of orientation conditioned key-point selection and localized feature extraction modules is used in parallel to supplement the features from the first path. By using orientation as a conditioning factor for adaptive key-point selection, the model learns to focus on the most informative parts of the vehicle. Additionally, we develop a fully convolutional two-stage key-point detection model inspired by the works of Kumar *et al.* [9] and Bulat *et al.* [2] for facial key-point detection and human pose estimation respectively.

The detailed architectures of each module in the proposed method are discussed in section 3. Through extensive experimentation, we show that the proposed Adaptive Attention model for Vehicle Re-identification (AAVER) approach improves the re-id accuracy on challenging datasets such as VeRi-776 [11, 12] and VehicleID [10]. In addition, the proposed vehicle key-point detection model, improves the accuracy by more than 7% over state of the art.

2. Related Work

In this section, we briefly review recent relevant works in the field of vehicle classification and re-identification. Learning a discriminating representation, requires a large-scale annotated data for training, especially for recent DCNN approaches. Yang *et al.* [27] released a large-scale car dataset (CompCars) for fine-grained vehicle model classification which consists of 1,687 car models and 214,345 images. The VehicleID dataset by Liu *et al.* [10] consists of 200,000 images of about 26,000 vehicles. In addition, Liu *et al.* [12, 13] published a high-quality multi-view vehicle re-id (VeRi-776) dataset. Yan *et al.* [26] released two high-quality and well-annotated vehicle datasets, namely VD1 and VD2, with diverse annotated attributes, containing 1,097,649 and 807,260 vehicle images captured in different cities.

Moreover, besides datasets for training, Tang *et al.* [22] claimed traditional hand-crafted features are complementary to deep features and thus fused both features to realize an improved representation. Instead, Cui *et al.* [4] fused features from various DCNNs trained with different objectives. Furthermore, Liu *et al.* [12, 13] used multi-modal features, including visual features, license plate, camera location, and other contextual information, in a coarse-to-fine vehicle retrieval framework. To augment the training data for robust training, [24] used a generative adversarial network to synthesize vehicle images with diverse orientation and appearance variations. [31] learns a viewpoint-aware representation for vehicle re-id through adverserial learning and a viewpoint-aware attention model.

Besides global features, Liu *et al.* [14] extracted discriminative local features from a series of local regions of a vehicle by a region-aware deep model. Different from these approaches, the proposed method leverages orientation to adaptively select the regions of attentions.

Another effective strategy to learn the discriminative representation is metric learning. Zhang *et al.* [29] proposed an improved triplet loss which performs joint optimization with an auxiliary classification loss as a regularizer in order to characterize intra-sample variances. Bai *et al.* [1] introduced Group-Sensitive triplet embeding to better model the intra-class variance. Shen *et al.* [20] also proposed to improve the matching performance by making use of spatio-temporal information; they developed a Siamese-CNN with path LSTM model which generates the corresponding candidate visual-spatio-temporal paths of an actual vehicle image by a chain-based Markov random field (MRF) model with a deeply learned potential function. In contrast, the proposed method uses the L_2 softmax [19] loss function as it has shown impressive performance for the task of face verification and trains faster compared to triplet loss-based methods such as [29] without the hassle of sampling hard triplets.

3. Adaptive Attention Vehicle Re-identification (AAVER)

The entire pipeline of the proposed method AAVER is composed of three main modules: **Global Feature Extraction**, **Vehicle Key-Point and Viewpoint Estimation**, and **Adaptive Key-Point Selection and Feature Extraction** which is followed by a re-ranking based post-processing. Figure 2 shows the diagrammatic overview of our method.

In AAVER, the global feature extraction module is responsible for extracting the macroscopic features (f_g) of the vehicles. By looking at the entire vehicle, this model tries to maximally separate the identities in the feature space. However, this model may fail to take into account subtle differences between similar cars, most extremely the ones that are of the same make, model and color. Therefore, the features generated by this module are supplemented with features (f_l) from the localized feature extraction module. This can be achieved by the adaptive attention strategy using the proposed key-point and orientation estimation network.

In order to estimate the vehicle key-points, we draw in-

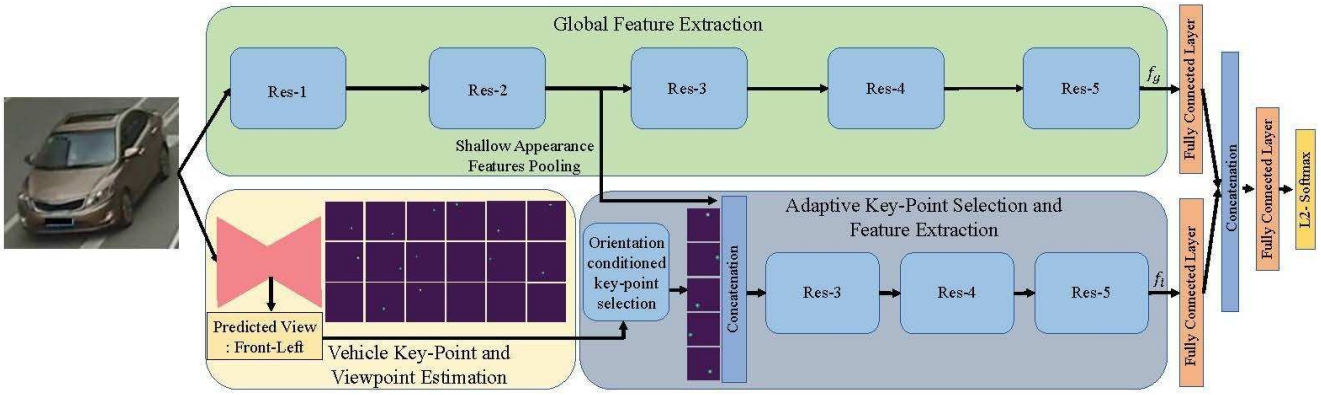


Figure 2: Adaptive Attention Vehicle Re-identification (AAVER) Model Pipeline. The input vehicle image is processed in parallel along two paths: In the first path, the global appearance features (f_g) are extracted. The second path is responsible for detecting vehicle key-points and predicting its orientation, after which localized features (f_l) are extracted based on adaptive key-point selection. Subsequently, the two feature vectors f_g and f_l are fused with a shallow multi-layer perceptron.

spiration from literature on facial key-point detection and human pose estimation. Inspired by [2,9] we employ a two-stage model to predict the vehicle's orientation and landmarks in a coarse to fine manner; the coarse heatmaps predicted by a DCNN are refined using a shallower network.

Finally, we use the proposed adaptive key-point selection module to select a subset of most informative key-points and pool features from early layers of the global feature extraction module to extract localized features around the selected key-points. The features obtained from the two paths of AAVER are then merged using a multi-layer perceptron. The entire model can be trained end-to-end using any differentiable loss function. In our work, we use the L_2 softmax loss as proposed in [19]. During inference, we use the features from penultimate fully connected layer as the representation of a given vehicle. Additionally, we also perform re-ranking [30] as a post processing step.

Each module is described in detail in the following sub sections. Pytorch deep learning framework [17] has been used in all of the experiments.

3.1. Global Feature Extraction

For extracting the global appearance features, we employ ResNet-50 and ResNet-101 [6] as backbone networks and also adopt them as our baseline models. We initialized the weights of these models using the weights from the models pre-trained on the CompCars dataset. A 2048-dimensional features vector from the last convolutional layer of ResNet is then fed to a shallow multi-layer perceptron. This network is trained using the L_2 softmax loss function which constrains the feature vectors extracted by the network to lie on a hyper-sphere of radius α. This enables the network to embed features of identical vehicles together while pushing apart the features from different vehicles. It is mathematically expressed as:

$$\mathcal{L}_S = -\log \frac{\exp(\mathbf{W}_y^T(\frac{\alpha \mathbf{x}}{\|\mathbf{x}\|_2}) + b_y)}{\sum_{j=1}^{N} \exp(\mathbf{W}_j^T(\frac{\alpha \mathbf{x}}{\|\mathbf{x}\|_2}) + b_j)} \quad (1)$$

where $\mathbf{x}$ is the feature vector corresponding to class label y, $\mathbf{W}_j$ is the weight and b_j is the bias corresponding to class j, α is a positive trainable scalar parameter, and N is the number of classes respectively.

3.2. Vehicle Key-Point and Orientation Estimation

In this work, a two-stage model is proposed for key-point estimation. In the first stage, a VGG-16 [21] -based fully convolutional network is employed to do a coarse estimation of the location of N_1 ($N_1 = 21 = 20$ key-points plus background) heatmaps of size $H \times W$ (56×56). This network is trained using a per-pixel multi-class cross entropy loss defined as follows:

$$\mathcal{L}_1 = \frac{-1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \log(\frac{\exp(\mathbf{x}_{i,j}(t^*{}_{i,j}))}{\sum_{k=1}^{N_1} \exp(\mathbf{x}_{i,j}(k))}) \quad (2)$$

where $\mathbf{x}_{i,j}$ is the vector corresponding to pixel location i and j across all output channels and $t^*_{i,j}$ is the ground-truth class label for that pixel location. After training the first stage, the weights of this network are frozen for training of the subsequent stage. The left side of Figure 3 depicts the the output of the first stage for a sample vehicle image.

Although the responses of the first stage can be used for the prediction of visible key-point locations, there might be erroneous activations in the heatmaps that correspond to invisible key-points. Consequently, we use the second

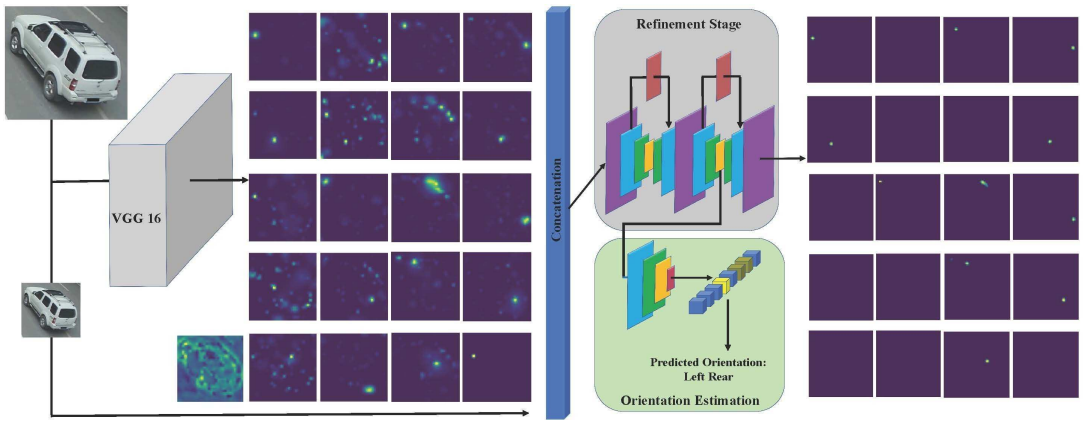


Figure 3: Vehicle key-point and orientation estimator network. VGG 16 network outputs 21 coarse heatmaps corresponding to the 20 vehicle landmarks and the background (Response maps on the left). A two-stack hourglass network refines 20 key-points heatmaps (response maps on the right) excluding background channel and predicts the vehicle's orientation.

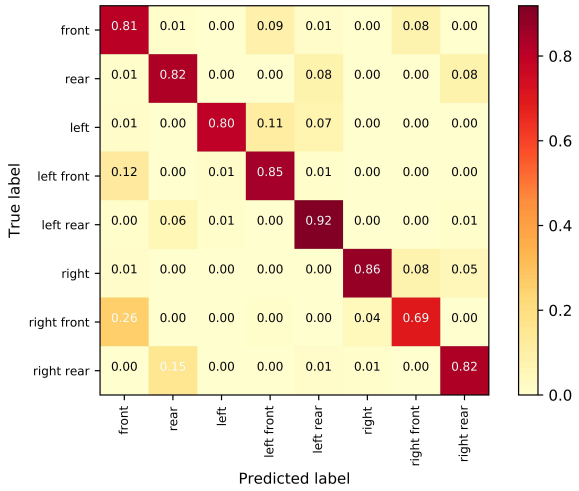


Figure 4: The confusion matrix of the vehicle orientation estimation network

stage that takes in the sub-sampled version of the input image and the coarse estimates of key-points to refine the results. The refinement network follows the hourglass architecture introduced in [16] which is commonly used for refining heatmaps and reducing artifacts due to invisible key-points. In the second stage, coarse heatmaps estimated from the first stage, are refined through a two-stack hourglass network with skip connections. Along with refining the estimated key-points, the orientation of the vehicle is also predicted through a parallel branch composed of two fully connected layers designed to classify the orientation into eight classes as defined in [23]. This multi-task learning helps the refinement network to make accurate predictions of the visible key-points while reducing the response of invisible key-points. Figure 3 shows the overall schematic flow of the two-stage network.

To train the heatmap refinement and orientation branches we use Mean Square Error (MSE) and cross entropy loss respectively. Equation 3 represents the loss function used for the second stage. It is worth mentioning that in the second stage we are only interested in foreground heatmaps, hence, we exclude the refinement of the background channel.

$$\mathcal{L}_2 = \mathcal{L}_H + \lambda * \mathcal{L}_O \tag{3}$$

where $\mathcal{L}_H$ is the heatmap regression loss:

$$\mathcal{L}_H = \sum_{k=1}^{N_2} \sum_{i=1}^{H} \sum_{j=1}^{W} | \, h_k(i,j) - h^*{}_k(i,j) \, |^2 \tag{4}$$

and $\mathcal{L}_O$ is the orientation classification loss:

$$\mathcal{L}_O = -\log\left(\frac{\exp(\mathbf{p}(p^*))}{\sum_{i=1}^{N_\mathbf{p}} \exp(\mathbf{p}(i))}\right). \tag{5}$$

In Equation (4), $N_2 = N_1 - 1$, $h_k(i,j)$ and $h_k^*(i,j)$ are predicted and ground-truth heatmaps in stage 2 for the k^{th} key-point at locations i and j respectively. $\mathbf{p}$, p^* and N_p in Equation (5) constitute the predicted orientation vector, the corresponding ground-truth orientation and number of classes respectively. Finally, λ in Equation (3) is a weight to balance the losses used in model optimization. In our experiments, λ is set to 10 obtained after cross-validation. In the right hand side of Figure 3 which shows the the output of the second stage, it can be observed that the initial coarse estimates of key-points have been refined.

3.3. Adaptive Key-Point Selection and Feature Extraction

Subtle differences in similar vehicles mostly occur close to vehicle landmarks, *e.g.* same car make and models of

Table 1: Seven Prominent key-points in each orientation group

Orientation Group	Visible Key-Points
Front	[11, 12, 7, 8, 9, 13, 14]
Rear	[18, 16, 15, 19, 17, 11, 12]
Left	[8, 1, 11, 14, 15, 2, 17]
Left Front	[9, 14, 6, 8, 11, 1, 15]
Left Rear	[2, 17, 15, 11, 14, 19, 1]
Right	[7, 3, 12, 13, 16, 4, 18]
Right Front	[9, 13, 5, 7, 12, 3, 16]
Right Rear	[3, 4, 12, 16, 18, 19, 13]

same color might be distinguishable through their window stickers, rims, indicator lights on the side mirrors, etc. This can be achieved by focusing the attention on parts of the image that encompasses these distinctions. To this end, regions of interest within the image are identified based on the orientation of the vehicle; after which features from the shallower layer of the global appearance model are pooled. As suggested in [28] these pooled features contain contextual rather than abstract information. Later, deep blocks (Res3, Res4 and Res5) of another ResNet model are used to extract supplementary features corresponding to the regions of interest.

In [23], vehicle's orientation is annotated into eight different classes, *i.e.* rear, left ,left front, left rear, right, right front and right rear; however, there is no absolute boundary between two adjacent orientations. For instance, for the case of right and right front, the network gets confused between the two classes when trained for orientation prediction; this can be observed in Figure 4 which shows the confusion matrix for the eight-class classification problem. To overcome this issue, we designed a key-point selector module that takes the predicted orientation likelihood vector and adaptively selects the key-points based on the likelihoods.

In order to achieve this, we constructed eight groups shown in Table 1 corresponding to each of the eight orientations of a vehicle and its two adjacent orientations. During inference, the likelihood of each orientation group is calculated and the one with the highest probability is picked. Also, experimentally it was observed that for each orientation group at least seven key-points are always visible. Consequently, given the orientation group with the highest probability we select the seven heatmaps shown in Table 1 corresponding to the respective orientation group. These orientation groups are named based on their center orientation *e.g.* the group that contains left front, front and right front is named front.

After obtaining the seven heatmaps, for each map, a Gaussian kernel with $\sigma = 2$ is placed in the location of the map's peak, *i.e.* the key-point location. This is done in order to emphasize the importance of the surrounding areas around the key-points as they may have discriminative information.

Following the adaptive heatmap selection and dilation by the Gaussian kernel, is the localized feature extraction (f_l) by Res3, Res4 and Res5 blocks of the parallel ResNet model. The input to this sub-network is the concatenation of the seven dilated heatmaps of shape $7 \times 56 \times 56$ and the pooled global features of shape $256 \times 56 \times 56$. Finally, the localized features f_l is concatenated with the global appearance features f_g and passed through a multi-layer perceptron followed by L_2 softmax loss function (refer to Figure 2). Given that features are normalized, we use cosine similarity to calculate the similarity score between image pairs.

3.4. Post Processing Step: Re-Ranking

In general, Re-ID can be regarded as a retrieval problem. Given a probe vehicle, we want to search in the gallery for images containing the same vehicle in a cross-camera mode. After an initial ranking list is obtained, a good practice consists of adding a re-ranking step, with the expectation that the relevant images will receive higher ranks. Such re-ranking steps have been mostly studied in generic instance retrievals such as [18], [3], [7] and [30]. The main advantage of many re-ranking methods is that they can be implemented without requiring additional training samples, and also can be applied to any initial ranking list.

Significant amount of research in person re-id goes into re-ranking strategies and vehicle re-id is lacking in that aspect. Most of the state of the art methods for vehicle re-id do not perform re-ranking on their initial ranking list. We use the re-ranking strategy proposed by Zhong et al. [30] in our work.

4. Experiments

Here we first present the two large-scale datasets used for the vehicle re-identification task and their evaluation protocols, after which we describe the implementation details of the proposed method.

4.1. Datasets

To the best our knowledge, there are mainly two large scale vehicle datasets that are publicly available and are designed for the task of vehicle re-identification: VeRi-776 [11], [12] and VehicleID [10].

VeRi-776 dataset consists of 49,357 images of 776 distinct vehicles that were captured with 20 non-overlapping cameras in variety of orientations and lighting conditions. Out of these images, 37,778 (576 identities) and 11,579 (200 identities) have been assigned to training and testing respectively. For the query set, 1,678 images have been selected from the testing set. The evaluation protocol for this

dataset is as follows: for each probe image in the query set the corresponding identity and the camera ID from which the image is captured is gathered. The gallery is constructed by selecting all the images in the testing set except the ones that share the same identity and camera ID as the probe. Evaluation metrics adopted for this dataset are mean Average Precision (mAP), Cumulative Match Curve (CMC) for top 1 (CMC@1) and top 5 (CMC@5) matches.

VehicleID is another large-scale dataset used for vehicle retrieval task and is composed of $221,567$ images from $26,328$ unique vehicles. Half of the identities, *i.e.* $13,164$, are reserved for training while the other half are dedicated for evaluation. There are 6 test splits for gallery sizes of 800, 1600, 2400, 3200, 6000 and $13,164$. In the recent works [20, 23] the first three splits have been used. The proposed evaluation protocol for each split in VehicleID dataset is to randomly select an image for each of the identities to form the gallery of respective size and use the rest of the images for query. This procedure is repeated ten times and the averaged metrics, CMC@1 and CMC@5, are reported.

4.2. Implementation Details

In our implementation, all the input images were resized to $(224, 224)$ and normalized by the ImageNet dataset [5] mean and standard deviation. Also, in all of our experiments we used batch training with size of 150 and Adam optimizer [8] with the learning rate of $1e-4$.

Initially, we fine-tuned our baseline models (see section 3.1) on VeRi-776 and VehicleID datasets separately, for 20 epochs. Then, we initialized the key-point and orientation estimation network with ImageNet pre-trained weights. The first stage of this network was trained for 40 epochs; afterwards the second stage was trained for 40 epochs as well.

Next, we trained the orientation conditioned feature extraction branch for each of VeRi-776 and VehicleID datasets for 20 epochs. Finally, we select the network's output of the penultimate layer as the feature vector corresponding to the input vehicle image.

5. Experimental Evaluations

We first present the evaluation results of our vehicle key-point and orientation estimation model followed by the evaluation of the proposed method AAVER on both VeRi-776 and VehicleID datasets.

5.1. Vehicle Key-Point and Orientation Estimation Evaluation

In order to evaluate the performance of the proposed two-stage key-point detection model, we use the Mean Square Error (MSE) in terms of pixels for the location of visible key-points in 56×56 maps over the test set of VeRi-776 key-point dataset. Table 2 shows the MSE of our model after first and second stages. Moreover, we measured the accuracy of the model for viewpoint classification. It can be observed that the refinement stage reduces the key-point localization error by 20% compared to the first stage.

To the best of our knowledge, [23] is the only work on the VeRi-776 key-point and orientation estimation dataset. [23] used the averaged distance between estimated and ground-truth locations of all visible key-points for evaluation. If the distance is less than a threshold value (r_0 in terms of pixels in 48×48 map), the estimation is considered to be correct. We follow the same protocol to compare the precision with [23] and Table 2 shows the result of this comparison.

Table 2: Accuracy evaluation and comparison of the vehicle landmark and orientation estimation network

	Stage 1	Stage 2
Key-point localization MSE (pixel)	1.95	**1.56**
Orientation Accuracy	-	84.44%
Key-Point Precision Comparison		
Model	$r_0 = 3$	$r_0 = 5$
OIFE [23]	88.8%	92.05%
Ours	**95.30%**	**97.11%**

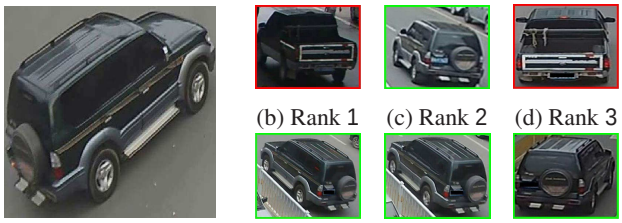
5.2. Evaluation Results on VeRi-776

Table 3 summarizes the results of the global appearance model (baseline) and the proposed AAVER model with adaptive attention. Note that in both ResNet-50 and ResNet-101 -based architectures, there is a significant improvement in mAP and CMC@1 scores after incorporating adaptive attention. This indicates that conditioning on the orientation of the vehicle and selecting corresponding key-points enables the network to focus more on parts that contains minute differences in similar cars. This claim is further studied in section 5.5. Unsurprisingly, we also observe that ResNet-101 shows better performance compared to ResNet-50 under similar settings.

Table 3: Performance comparison between baseline and the proposed method on VeRi-776 dataset

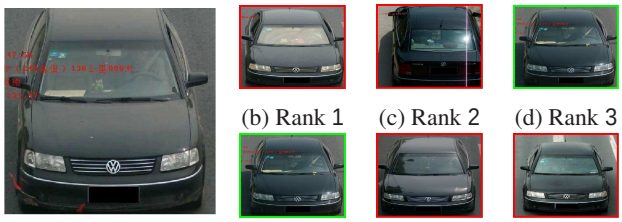
Model		mAP	CMC@1	CMC@5
Baseline	ResNet-50	52.88	83.49	92.31
	ResNet-101	55.75	84.74	94.34
AAVER	ResNet-50	58.52	88.68	94.10
	ResNet-101	**61.18**	**88.97**	**94.70**

Figure 5 plots the probe image and the top three returns of each baseline and the proposed model. It can be observed that AAVER significantly improves the performance over the baseline.



(a) Probe Image (b) Rank 1 (c) Rank 2 (d) Rank 3 (e) Rank 1 (f) Rank 2 (g) Rank 3

Figure 5: Top three returned results of the baseline model (sub-figures b-d) versus the AAVER model (sub-figures e-g) on VeRi-776 dataset



(a) Probe Image (b) Rank 1 (c) Rank 2 (d) Rank 3 (e) Rank1 (f) Rank 2 (g) Rank 3

Figure 6: Top three returned results of the baseline model (sub-figures b-d) versus the AAVER model (sub-figures e-g) on VehicleID dataset

5.3. Evaluation Results on VehicleID

Images in this dataset have less variations in viewpoint, *i.e.* mostly front and rear, compared to VeRi-776 dataset. For this dataset, the evaluation metrics are only CMC@1 and CMC@5 as there is only one true match in the gallery for each probe image. Table 4 presents the re-identification results of baseline and the proposed models over test splits. As compared to baseline models, a significant increase in performance is observed when features from adaptive attention-based path are fused with global appearance features.

Table 4: Performance comparison between baseline and proposed method on VehicleID dataset

		Baseline Model		AAVER Model	
	Split	ResNet-50	ResNet-101	ResNet-50	ResNet-101
CMC@1	800	67.27	70.03	72.47	**74.69**
	1600	62.03	65.26	66.85	**68.62**
	2400	55.12	59.04	60.23	**63.54**
CMC@5	800	89.05	89.81	93.22	**93.82**
	1600	84.31	84.96	89.39	**89.95**
	2400	80.04	80.60	84.85	**85.64**

Figure 6 shows an examples of a query from VehicleID dataset and the top three results returned by both global and adaptive attention model.

5.4. Comparison with State of the Art Methods

In this section, we compare AAVER model with ResNet-101 backbone against the recent state of the art methods. The results of this comparison are presented in Table 5.

From Table 5, it can be observed that our proposed method is among the top performers of vehicle re-identification task and is the state of the art for most of the evaluation metrics on both VeRi-776 and VehicleID datasets. Note that in the absence of a deterministic test set for VehicleID dataset, one cannot provide the basis for a fair comparison. The reason lies in the fact that random gallery construction yields different evaluation results with relatively high variance even when averaged over ten repetitions. Finally, we have to emphasize on the necessity of using re-ranking as a post processing step whenever there are multiple instances of the probe image in the gallery. Here for the VeRi-776 dataset, re-ranking shows significant improvement and results in state of the art mAP and CMC@1 scores. Note that for VehicleID dataset re-ranking is not applicable as there is only one true match in the gallery for each probe image.

5.5. Ablation Studies

We designed a set of experiments to study the impact of complementary information that the orientation conditioned branch provides. Note that in these experiments we only use the test split 800 for the VehicleID dataset To this end, the following experiments have been conducted:

1. In the first experiment we examined the depth of the layer in the global branch from which the global features are pooled and then fed to the orientation conditioned branch. To investigate this we tried pooling features after Res2, Res3 and Res4 blocks of spatial size of 56×56, 28×28 and 14×14. Table 6 demonstrates the results of this experiment. It can be observed that as we go from shallow to deeper layers, the features become more abstract and focusing on parts of deep feature maps do not help in providing a robust representation of vehicles with minute differences.

2. In our method, we use two streams for extracting global and local features from vehicle images, so we were keen to see whether a single branch can extract discriminative features that encompass global as well as local differences. To test this hypothesis, instead of pooling features from the global branch we fused the selected heatmaps into the global branch by concatenation and used the output as the representation for a vehicle image. Table 7 depicts the result of this experiment, for both VeRi and VehicleID datasets. we can infer that the re-identification performance drops significantly by relying on a single branch.

Table 5: Comparison with recent methods and state of the arts

	Dataset								
	VeRi-776			VehicleID					
				Test size = 800		Test size = 1600		Test size = 2400	
Method	mAP	CMC@1	CMC@5	CMC@1	CMC@5	CMC@1	CMC@5	CMC@1	CMC@5
SCPL [20]	58.27	83.49	90.04	-	-	-	-	-	-
OIFE [23]	48.00	65.9	87.7	-	-	-	-	-	-
VAMI [31]	50.13	77.03	90.82	63.12	83.25	52.87	75.12	47.34	70.29
RAM [14]	61.5	88.6	94.0	**75.2**	91.5	**72.3**	87.0	**67.7**	84.5
AAVER	61.18	88.97	**94.70**	74.69	**93.82**	68.62	**89.95**	63.54	**85.64**
AAVER + Re-ranking	**66.35**	**90.17**	94.34	-	-	-	-	-	-

Table 6: Experiment 1: Depth of pooled global features

Dataset	features size	mAP	CMC@1	CMC@5
VeRi-776	56×56	**0.612**	**88.97**	**94.70**
	28×28	0.608	88.50	94.58
	14×14	0.597	85.88	93.03
VehicleID	56×56	-	**74.69**	**93.82**
	28×28	-	72.60	93.24
	14×14	-	71.09	92.13

Table 7: Experiment 2: Single versus Dual-branch feature extraction

Dataset	Type	mAP	CMC@1	CMC@5
VeRi-776	Single	0.528	80.93	90.52
	Dual	**0.612**	**88.97**	**94.70**
VehicleID	Single	-	69.61	91.45
	Dual	-	**74.69**	**93.82**

3. In the final set of experiments we scrutinize the way in which the information from vehicle key-point heatmaps are incorporated in the proposed model. Our work is in some aspects similar to [23] which groups a fixed set of key-points and combines all the corresponding heatmaps into one map by adding them together. Therefore, we conduct this experiment under the same settings as of [23]. Table 8 shows the results of these experiments. The type "Combined" in Table 8 refers to the method in [23]. We can conclude that using all heatmaps combined into one group does not result in competitive results as the adaptive selection of heatmaps. This validates the hypothesis that not all the key-points contribute to a discriminative representation of the vehicle.

6. Conclusions and Future Work

In this paper, we present a robust end-to-end framework for state of the art vehicle re-identification. We present a dual path model AAVER which combines macroscopic global features with localized discriminative features to efficiently identify a probe image in a gallery of varying sizes. In addition, we establish benchmarks for key-point detection and orientation prediction on VeRi-776 dataset. Lastly, we advocate for the adoption of re-ranking when considering the performance of future vehicle re-identification methods. Adaptive key-point selection conditioned on vehicle orientation is vital for discriminating between vehicles of the same make, model and color. Evaluating on both VeRi-776 and VehicleID shows the strength of our proposed method. Lastly, we conduct an ablation study to understand the influence of the adaptive key-point selection step.

In the future, we plan to extend our key-point module to align vehicle images to a canonical coordinates system before comparing a given pair of images. Similarly, we can learn a 3D representation of vehicles to be used in other tasks such as vehicular speed estimation.

Table 8: Experiment 3: Key-points heatmaps utilization

Dataset	Type	mAP	CMC@1	CMC@5
VeRi-776	Combined [23]	0.606	87.66	94.17
	AAVER	**0.612**	**88.97**	**94.70**
VehicleID	Combined [23]	-	71.79	92.10
	AAVER	-	**74.69**	**93.82**

7. Acknowledgement

This research is supported in part by the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative, It is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 20(9):2385–2399, Sep. 2018. 2

[2] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 2, 3

[3] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 5

[4] C. Cui, N. Sang, C. Gao, and L. Zou. Vehicle re-identification by fusing multiple deep neural networks. In *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2017. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009. 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[7] Herve Jegou, Hedi Harzallah, and Cordelia Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 5

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[9] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3

[10] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016. 2, 5

[11] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re-identification in urban surveillance videos. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016. 2, 5

[12] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European Conference on Computer Vision*, pages 869–884. Springer, 2016. 2, 5

[13] X. Liu, W. Liu, T. Mei, and H. Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018. 2

[14] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. Ram: a region-aware deep model for vehicle re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 2, 8

[15] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. A large-scale dataset for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 10

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 4

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 3

[18] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR 2011*, pages 777–784. IEEE, 2011. 5

[19] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 2, 3

[20] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1918–1927, 2017. 2, 6, 8

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[22] Y. Tang, D. Wu, Z. Jin, W. Zou, and X. Li. Multi-modal metric learning for vehicle re-identification in traffic surveillance environment. In *IEEE International Conference on Image Processing (ICIP)*, pages 2254–2258, 2017. 2

[23] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 379–387, 2017. 1, 2, 4, 5, 6, 8

[24] F. Wu, S. Yan, J. S. Smith, and B. Zhang. Joint semi-supervised learning and re-ranking for vehicle re-identification. In *IEEE Conference on Pattern Recognition (ICPR)*, 2018. 2

[25] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[26] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

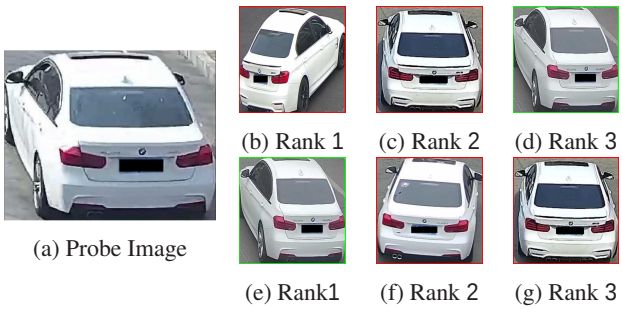
[27] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3973–3981, 2015. 2

[28] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. 5

[29] Y. Zhang, D. Liu, and Z.-J. Zha. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1386–1391, 2017. 2

[30] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 3, 5

[31] Y Zhou and L Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn*, volume 2, 2018. 2, 8



(a) Probe Image (b) Rank 1 (c) Rank 2 (d) Rank 3 (e) Rank1 (f) Rank 2 (g) Rank 3

Figure 7: Top three returned results of the baseline model (sub-figures b-d) versus the AAVER model (sub-figures e-g) on VeRi-Wild dataset

Supplementary Material

In this section we present the performance of our method on the newly released VeRi-Wild dataset [15]. This dataset is collected in a network of 174 surveillance cameras covering a large urban area and captures unconstrained scenarios. It is composed of 416,314 images (277,797/138,517 for train/test sets) of 40,671 different vehicle identities. The test set, similar to VehicleID dataset is divided into three Small (41,861), Medium (69,389) and Large (138,517) images splits. The evaluation for this dataset follows the same protocol as VeRi-776. Table 9 summarizes the result of our baseline and proposed AAVER model.

Table 9: Performance comparison between baseline and proposed method on VeRi-Wild dataset

		Baseline Model		AAVER Model	
	Split	ResNet-50	ResNet-101	ResNet-50	ResNet-101
mAP	Small	60.22	60.99	60.62	**62.23**
	Medium	51.21	52.49	51.77	**53.66**
	Large	38.89	38.99	40.42	**41.68**
CMC@1	Small	71.37	72.97	74.60	**75.80**
	Medium	62.84	65.02	65.76	**68.24**
	Large	52.00	54.65	56.03	**58.69**
CMC@5	Small	90.10	91.57	91.60	**92.70**
	Medium	86.58	87.30	87.16	**88.88**
	Large	77.48	79.52	79.67	**81.59**

From Table 9 it can be seen that for all splits of VeRi-Wild dataset like VeRi-776 and VehicleID datasets, a significant boost is obtained by conditioning the features on the vehicle's orientation and corresponding key-points.

Figure 7 shows an examples of a query from VeRi-Wild dataset and the top three results returned by both global and adaptive attention models.