# Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification

Zhongdao Wang[1,2], Luming Tang[1,2], Xihui Liu[1,2], Zhuliang Yao[1,2], Shuai Yi[1], Jing Shao[1], Junjie Yan[1], Shengjin Wang[2], Hongsheng Li[3], Xiaogang Wang[3]
[1]SenseTime Group Limited [2]Tsinghua University [3]The Chinese University of Hong Kong
yishuai@sensetime.com hsli@ee.cuhk.edu.hk

## Abstract

*In this paper, we tackle the vehicle Re-identification (ReID) problem which is of great importance in urban surveillance and can be used for multiple applications. In our vehicle ReID framework, an orientation invariant feature embedding module and a spatial-temporal regularization module are proposed. With orientation invariant feature embedding, local region features of different orientations can be extracted based on 20 key point locations and can be well aligned and combined. With spatial-temporal regularization, the log-normal distribution is adopted to model the spatial-temporal constraints and the retrieval results can be refined. Experiments are conducted on public vehicle ReID datasets and our proposed method achieves state-of-the-art performance. Investigations of the proposed framework is conducted, including the landmark regressor and comparisons with attention mechanism. Both the orientation invariant feature embedding and the spatio-temporal regularization achieve considerable improvements.* [1] [2] [3]

## 1. Introduction

In this paper, we target on the problem of vehicle re-identification (ReID), which aims to identify all the images of the same vehicle from a large gallery database. Such a task is particularly useful when the car licence plate is occluded or cannot be seen clearly. Vehicle ReID methods can be used in these scenarios to effectively locate vehicles of interest from surveillance databases. They have extensive applications in intelligent surveillance and attract increasing attention in recent years. Compared with the problem of person ReID, which has been studied for years, vehicle ReID is a recently proposed research topic. There exist specific characteristics and challenges for this problem.

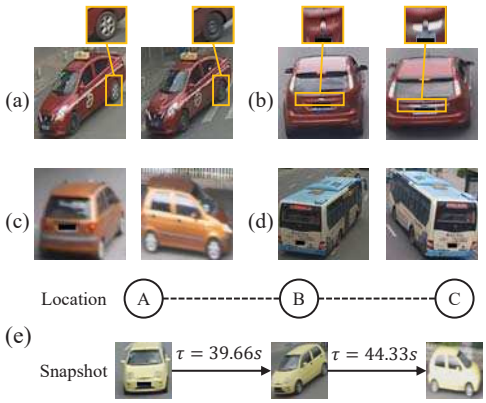


Figure 1. Difficulties of the problem of vehicle ReID. (a-b) Vehicle image pairs that share quite similar overall appearances. They can only be distinguished from local regions like the wheels in (a) and the logos in (b). (c-d) Different faces of one vehicle may have different visibility, which results in difficulties in aligning the features of different faces. In our proposed orientation invariant network, visible faces such as the right faces in (c) and the back faces in (d) are assigned with a larger weight in the final feature embedding process. (e) Spatio-temporal constraints of a vehicle's appearance should be satisfied, *e.g.* around 39.66 seconds between A and B and 44.33 seconds between B and C.

Firstly, some specific regions in vehicle images are important for vehicle ReID. Different from person images, which contain rich textures, the vehicles are generally solid-colored and sometimes the color patterns between different vehicles can be quite similar. For example, as shown in Fig. 1 (a-b), the cars are all red and are difficult to distinguish based on their general appearances. The wheel regions in (a) and the logo regions in (b) are keys to determine whether those vehicles are the same ones. Most existing

[1]Z. Wang and L. Tang share equal contribution.
[2]S. Yi and H. Li are the corresponding authors.
[3]This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14213616, CUHK14206114, CUHK14205615, CUHK419412, CUHK14203015, CUHK14239816, CUHK14207814, in part by the Hong Kong Innovation and Technology Support Programme Grant ITS/121/15FX, and in part by the China Postdoctoral Science Foundation under Grant 2014M552339.

vehicle ReID approaches [8, 9, 17] focus on the whole image and such subtle differences cannot be well taken into account. In our proposed method, region features are calculated based on 20 vehicle key point locations. In this way, vehicles with subtle differences can be well distinguished.

Secondly, there are always some key points not visible for vehicle images in one view. If we simply consider a vehicle as a cube with four sides, at most two of them could be visible at each time. As shown in Fig. 1 (c), the same car is captured in two views. The frontal face is invisible in (c1) while the left face is invisible in (c2). It is similar for the bus shown in Fig. 1 (d). Therefore, comparing whole-image features between vehicle images with different views is generally not optimal. In our framework, the 20 key points are clustered into four sets based on their orientations (front, back, left, and right), so that the key points in each set share the same visibility (all visible or all invisible). In order to distinguish the visible key point sets from the invisible ones, an orientation based feature calculation module is proposed in our framework. Different learnable weights are assigned to different key point sets according to the orientation of the input vehicle image. For instance, large weights are assigned to the right face in (c), as well as the back face in (d). In this way, the visible key points can contribute more to the final decision while the influences of invisible key points are weakened by lower weights. In addition, orientation invariant feature embedding is proposed to transform the weighted region features into the final orientation invariant feature vector.

Lastly, spatio-temporal constraints are also helpful for vehicle ReID. As shown in Fig. 1 (e), if a car is observed in camera (A), it is more likely to be observed in camera (B) with a time delay around 39.66 seconds. In the proposed approach, a conditional spatio-temporal distribution is modeled to regularize the final ReID results.

The proposed framework is evaluated on two standard vehicle ReID datasets, *i.e*. VeRi-776 [9] and VehicleID [8]. Our proposed framework outperforms state-of-the-art vehicle ReID methods. It achieves a mAP of $0.514$ on the Veri-776 [9] dataset, $86\%$ higher than the the best result ($0.277$) in literature [9]. For the VehicleID [8] dataset, a Top-1 accuracy of $67.0\%$ can be achieved, which is $75\%$ higher than the the best result ($38.2\%$) in literature [8].

The contribution of this work can be summarized as follows. 1) A deep learning framework is proposed for vehicle ReID, which contains four main components. The orientation-based region proposal module and feature extraction module are proposed to capture vehicles' region appearance information thus different vehicles showing similar overall appearances can be better distinguished. The orientation invariant feature aggregation module is proposed so that the region features of different views can be aligned and combined. The spatio-temporal regularization module is proposed to utilize spatio-temporal constraints to regularize the final ReID results. 2) The proposed framework is evaluated on two vehicle ReID datasets. Significant performance improvements over existing methods are achieved. 3) Ablation study of the proposed framework is conducted to investigate the effectiveness of its individual compoents, which includes investigations on the key point regressor and the comparisons with attention mechanism.

## 2. Related Work

**Re-identification (ReID)** is widely studied in computer vision which has various important applications. Most existing ReID methods focused on the person ReID problem, which aims to find target persons in a large gallery set given probe images. Many hand-crafted features are proposed to capture visual features for pedestrians [1,5,6,12,16,28]. Recently, CNN-based features [2,21,22,27] have also achieved great progress on person ReID.

**Vehicle ReID** is a newly proposed research topic and has not received much attention. Recent works on vehicle ReID mainly concentrate on building retrieval pipelines and benchmarks. Liu *et al*. [9] released a high-quality multi-viewed vehicle ReID dataset (named VeRi-776) with 776 vehicle identities, and proposed a progressive retrieval pipeline by combining vehicle appearance features, license plates, and spatio-temporal information. Another large surveillance-nature vehicle ReID dataset (VehicleID) is proposed by Liu *et al*. [8], which contains more than 20,000 identities. The Coupled Clusters Loss (CCL) is proposed for performance evaluation on this benchmark dataset. However, all these approaches on vehicle ReID utilize global appearance features of the input vehicle but do not focus on specific local discriminative regions.

**Fine-grained vehicle model classification** is relevant to vehicle ReID. Both tasks focus on learning discriminative feature representations for vehicle appearance. Yang *et al*. [23] published a large scale dataset (CompCars) for fine-grained vehicle model classification, which is the largest vehicle model dataset. Dominik *et al*. [25] and Jakub *et al*. [17] proposed to using 3D-boxes for aligning different vehicle faces and three visible faces are used for accurate feature extraction. However, this method may introduce ambiguities when the three visible faces are different. In our proposed method, the local feature extraction and aggregation modules is used to solve this issue.

**Object key point localization** has many important applications, *e.g*., face alignment [14] and human pose estimation [13, 20]. Key point-based face alignment is conducted in most face recognition frameworks [15, 18]. Locations of key points are helpful as the learned features can be well aligned by the key points. However, vehicle key points are not well studied in existing literature. Our proposed method shows that vehicle key points can guide the learning and

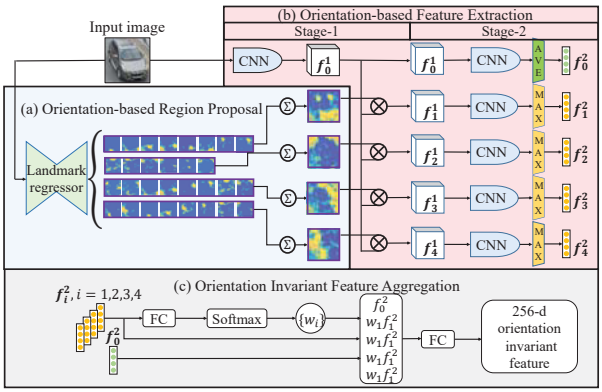


Figure 2. Illustration of the overall feature embedding pipeline, which consists of (a) the orientation-based region proposal module, (b) the orientation-based feature extraction module, and (c) the orientation-invariant feature aggregation module.

alignment of local regions in input vehicle images and improve the overall vehicle ReID performance.

## 3. Methodology

Our framework consists of two main components, the orientation invariant feature embedding component and the spatial-temporal regularization component. The pipeline of the orientation invariant feature embedding component is presented in Fig. 2, including three sub-modules, *i.e.* the orientation-based region proposal module (Sec. 3.1), the orientation-based feature extraction module (Sec. 3.2), and the orientation invariant feature aggregation module (Sec. 3.3). Firstly, vehicle images are fed into the region proposal module, which produces the response maps of 20 vehicle key points. The key points are then clustered into four orientation-based region proposal masks. Afterwards, the original image together with the four region proposal masks are utilized by the feature learning module to obtain one global feature vector and four region feature vectors. Finally, these features are fused by the aggregation module that outputs an orientation invariant feature vector. Besides learning the above mentioned appearance feature representations, a regularization strategy (Sec. 3.4) is adopted by modeling the spatio-temporal relations between the probe and gallery images. Training details of the proposed framework are introduced in Sec. 3.5.

### 3.1. Orientation-based Region Proposal

As shown in Fig. 2(a), a region proposal network is introduced in this section, which contains two steps, *i.e.* vehicle key point prediction and orientation-based region mask generation. The proposed region proposal network takes the image as input and estimates the vehicle key point locations. Four orientation-based region proposal masks are then generated based on the key points.

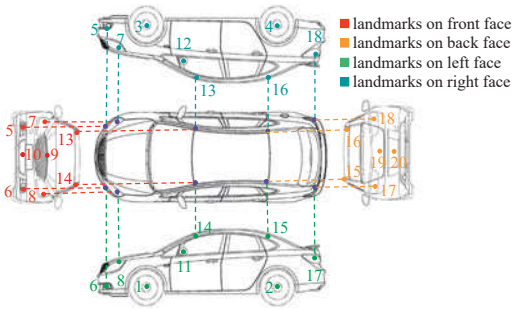


Figure 3. Illustration of the 20 selected vehicle key points. The 20 points are clustered into four sets based on their orientations, *i.e.*, the front face, the back face, the left face and the right face.

The first step of the region proposal network is to predict one response map for each vehicle key point. As listed in Table. 1 and shown in Fig. 3, 20 key points are specified for the vehicle ReID task. Instead of directly predicting boundary points or corner points, these key points are chosen as some discriminative locations or some main vehicle components, e.g. the wheels, the lamps, the logos, the rear-view mirrors, the license plates.

Inspired by the Stacked Hourglass Networks which generate response maps of human joints in a stacked coarse-to-fine manner for human pose estimation [13], an hourglass-like fully convolution network is adopted to generate vehicle key point response maps. The key point regressor takes the image as input and outputs one response map $F_i \in \mathbb{R}^{X \times Y}$ $(i \in 1, ..., 20)$ for each of the 20 key points, where $X$ and $Y$ are the horizontal and vertical dimensions of the feature maps.

The target response maps have Gaussian-like responses around the ground truth locations of key points and used as training supervisions. However, the Hourglass model [13] is computational expensive. Modifications to the network are made to reduce model complexity and also preserve the quality of output key point response maps. The input image size, the number of framework stages and the channel numbers of convolution layers are all reduced for fast computation. The per-pixel cross entropy loss between estimated response maps and the ground truth maps is adopted for training the network.

The second step of the region proposal network is to generate four orientation-based region masks. As introduced in Sec. 1, there are always some invisible regions for vehicles in specific orientations. To address the issue of invisible key points and make full use of the geometrical relationships among key points, the 20 key points indexed in Table 1 are assigned to four clusters, *i.e.*, $\mathcal{C}_1 = [5, 6, 7, 8, 9, 10, 13, 14]$, $\mathcal{C}_2 = [15, 16, 17, 18, 19, 20]$, $\mathcal{C}_3 = [1, 2, 6, 8, 11, 14, 15, 17]$, and $\mathcal{C}_4 = [3, 4, 5, 7, 12, 13, 16, 18]$, corresponding to the key points belonging to the vehicle's front face, back face, left

| | | | |
|---|---|---|---|
| 1 | left-front wheel | 11 | left rear-view mirror |
| 2 | left-back wheel | 12 | right rear-view mirror |
| 3 | right-front wheel | 13 | right-front corner of vehicle top |
| 4 | right-back wheel | 14 | left-front corner of vehicle top |
| 5 | right fog lamp | 15 | left-back corner of vehicle top |
| 6 | left fog lamp | 16 | right-back corner of vehicle top |
| 7 | right headlight | 17 | left rear lamp |
| 8 | left headlight | 18 | right rear lamp |
| 9 | front auto logo | 19 | rear auto logo |
| 10 | front license plate | 20 | rear license plate |

Table 1. Definition of the 20 selected vehicle key points.

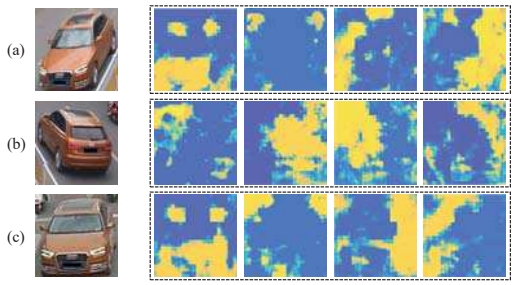


Figure 4. Examples of the four output response masks of the orientation based region proposal module. The input image and the corresponding four region masks, *i.e.* $\mathcal{R}_1$ of front face, $\mathcal{R}_2$ of the rear face, $\mathcal{R}_3$ of the left face and $\mathcal{R}_4$ of the right face, are shown in each row, from left to right respectively. Features of invisible faces, *e.g.*, $\mathcal{R}_2$ of (a) and (c), and $\mathcal{R}_1$ of (b), generally have low response masks. A feature aggregation method is adopted to reduce their impact on the final feature vector.

face, and right face, respectively. The final output region masks are computed as the summation of all the feature maps belonging to each cluster, *i.e.*

$$\mathcal{R}_i = \sum_{l \in \mathcal{C}_i} F_l, (i = 1, 2, 3, 4). \quad (1)$$

Examples of the output region masks $\mathcal{R}_i$ are shown in Fig. 4. From the results, we can observe that visible region masks generally have larger responses than the invisible ones, which demonstrate that the learned key point localization model not only estimates the key point locations but also discriminates the visible key points from the invisible ones. As the invisible region masks may not be suitable for feature extraction, the orientation invariant feature aggregation is proposed in Sec. 3.2 to handle such problem.

### 3.2. Orientation-based Feature Extraction

In the feature extraction module, deep convolutional neural network (CNN) is adopted to obtain one feature vector from the whole vehicle image and four orientation-related region feature vectors from the four corresponding regions. The network structure is shown in Fig. 2 (b), which contains two convolution stages, *i.e.* Stage-1, and Stage-2. The global feature and local features are extracted in a backbone-branch fashion.

In Stage-1, input images are resized to $192 \times 192$ and convolved by three convolution layers and two inception modules [19]. The output feature map is denoted as $f_0^{C1}$ with spatio size $12 \times 12$. In Stage-2, $f_0^1$ is assigned to five branches, including one global branch and four local region branches. For the global branch, the global feature map $f_0^1$ is convolved by one more inception module, and results in a set of $6 \times 6$ feature maps. Then global average pooling is applied on these feature maps to obtain a 1536-dimensional global feature vector $f_0^2$. For each local branch, the corresponding orientation-related region masks $\mathcal{R}_i (i = 1, 2, 3, 4)$ is resized to the same size as $f_0^1$, and $f_0^1$ is element-wisely multiplied by the region masks to obtain the local feature maps, *i.e.* $f_i^1 = f_0^1 \cdot \mathcal{R}_i (i = 1, 2, 3, 4)$. The results $f_i^1$ is further convolved by one more inception module. Global max pooling is adopted since the maximum responses are more suitable for guiding feature extraction from local regions. Every region branch outputs a 1536-dimensional feature vector $f_i^2 (i = 1, 2, 3, 4)$.

### 3.3. Orientation Invariant Feature Aggregation

As shown in Fig. 2(c), the feature aggregation module takes the five 1536-dimensional feature vectors, including one global feature $f_0^2$ and four local features $f_i^2, (i = 1, 2, 3, 4)$, as input and computes one 256-dimensional feature vector as output. In the aggregation module, the four local feature vectors are first concatenated and passed through a fully connected layer, yielding a set of scalars $\{e_i\}$. Then $\{e_i\}$ pass through the Softmax operator, producing a set of weights $\{w_i\}$, where $\sum_i w_i = 1, (i = 1, 2, 3, 4)$. The four local feature vectors are weighted by $\{w_i\}$ and concatenated together with the global feature vector $f_0^2$. The concatenation result $[f_0^2, w_1 f_1^2, w_2 f_2^2, w_3 f_3^2, w_4 f_4^2]^T$, is then fed into a fully connected layer and the output dimension is reduced to 256. The 256-dimensional feature vector is the final aggregated feature vector of the whole image, including the four local region features and one global region feature.

Examples in Fig. 5 demonstrates the effectiveness of the proposed orientation invariant feature aggregation module. Features of selected vehicle images in the VeRi-776 test set are projected to 2-dimensional space using t-SNE [10] and are visualized in Fig. 5(b). We can observe that features of the same identity can be clustered together, no matter which orientation the vehicle image is. Moreover, the input vehicle images and the corresponding learned weights of two clusters are shown in Figs. 5(a) and (c). For each image, the weights are learned for the four side faces, *i.e.* front, back, left, and right, and then the local features are fused based on these weights. We can observe that visible face are more likely to have higher weights than invisible ones.

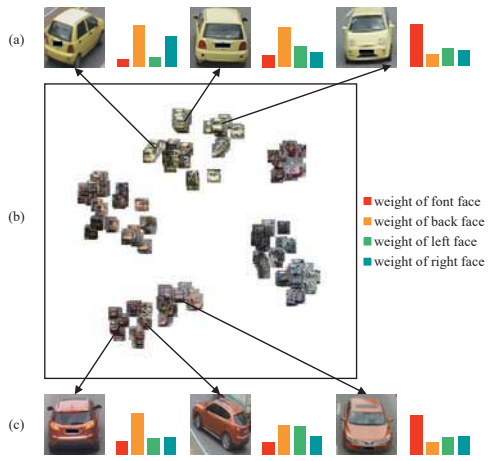


Figure 5. Illustration of the orientation invariant features with t-SNE [10]. (a,c) The input images of two different vehicles and their corresponding learned weights for different orientations, where visible faces are more likely to have higher weights. (b) 2D feature projections of selected vehicle images in the VeRi-776 test set using t-SNE.

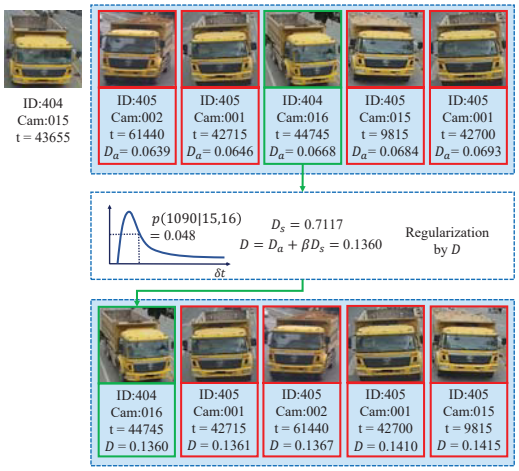


Figure 6. Illustration of the proposed spatio-temporal regularization step. Images in the first row are the query image and the top-5 retrieval results without spatio-temporal regularization. The green box represents the correct hit while red ones denote non-corresponding vehicles. The spatio-temporal distance $D_s$ between probe and gallery images are computed using the estimated log-normal distribution. The gallery images are regularized, and the results after regularization are shown in the bottom row.

### 3.4. Regularization by spatio-temporal Modeling

In real-world scenarios, appearance features may not be adequate enough to distinguish one vehicle from others, especially when the vehicles are of the same model without personalized decorations. However, in surveillance applications, the location and time information of a vehicle is easy to obtain. It is possible to refine vehicle search results with the help of such spatio-temporal information.

In order to investigate whether the spatio-temporal constraints are effective for vehicle ReID, we analyze the vehicle transition interval between pairs of cameras. For each camera pair, the transition interval can be modeled as a random variable that follows some probability distribution. Due to the Gaussian-like and long tail property of the transition interval, the logarithmic normal distribution is adopted to model this random variable. Given $l$ and $e$ as the leaving and entering cameras, the conditional probability of the transition interval $\tau$ between $l$ and $e$ can be estimated as the log-normal distribution $p(\tau|l,e)$,

$$\begin{aligned} p(\tau|l,e;\mu_{l,e},\sigma_{l,e}) &= \ln\mathcal{N}(\tau;\mu_{l,e},\sigma_{l,e}) \\ &= \frac{1}{\tau\sigma_{l,e}\sqrt{2\pi}}\exp\left[-\frac{(\ln\tau-\mu_{l,e})^2}{2\sigma_{l,e}^2}\right], \end{aligned} \tag{2}$$

where $\mu_{l,e} and \sigma_{l,e}$ are the parameters to be estimated for each camera pair $(l,e)$. The model parameters can be estimated by maximizing the following log-likelihood function,

$$L(\boldsymbol{\tau}|l,e;\mu_{l,e},\sigma_{l,e}) = \prod_{n=1}^{N}\left(\frac{1}{\tau_n}\right)\mathcal{N}(\ln\tau_n;\mu_{l,e},\sigma_{l,e}), \tag{3}$$

where $\tau_n \in \boldsymbol{\tau}(n = 1,2,3,...,N)$ is transition interval between camera pair $(l,e)$ sampled from the training set, and $\boldsymbol{\tau}$ contains all the time interval samples between the two cameras in the training set.

During the retrieval process, the appearance distance $D_a$ is first computed via the proposed orientation-invariant feature aggregation framework. The spatio-temporal distance $D_s$ is then computed for regularization. As shown in Fig. 6, the transition time interval between two cameras $(l,e)$ can be computed as $\tau = |t_l - t_e|$, where $t_l, t_e$ are the appearance time of this vehicle at these two cameras. The spatio-temporal probability can be computed as

$$p(\tau|l,e;\mu_{l,e},\sigma_{l,e}) = \ln\mathcal{N}(\tau;\mu_{l,e},\sigma_{l,e}). \tag{4}$$

High probabilities corresponds to small distances, thus

$$D_s = 1/(1+\boldsymbol{e}^{\alpha(p(\tau|l,e;\mu_{l,e},\sigma_{l,e})-0.5)}). \tag{5}$$

Finally the overall similarity distance between the probe and gallery images are calculated as the weighted summation,

$$D = D_a + \beta D_s, \tag{6}$$

where $\alpha$ is set to 2 and $\beta$ is set to 0.1 in our experiments.

### 3.5. Training Scheme

An alternative training strategy is adopted to train the proposed network, which include the following four steps.


Figure 7. Datasets used for training the proposed model, including (a) VeRi-776 [9], (b) VehicleID [8], (c) BoxCars21k [17],and (d) CompCars [23].

(i) The backbone of Stage-1 and the global branch of Stage-2 are trained from random initialization, by applying supervision to the global feature of full image region. (ii) With Stage-1 fixed, the four orientation branches are trained with parameters initialized as the global branch of Stage-2, since the global branch and the orientation branches in Stage-2 share the same structure. The four branches are trained separately by giving the classification label as supervision. (iii) With Stage-1 and all branches of Stage-2 fixed, the orientation invariant feature aggregation module is trained. (iv) Initializing all the modules with parameters learned from the above steps, and all the parameters are jointly fine-tuned. When training the model, existing vehicle datasets are used and the cross-entropy classification loss is adopted.

# 4. Experiments

## 4.1. Datasets

Four existing vehicle datasets are used to train the proposed orientation invariant network, including VeRi-776 [9], VehicleID [8], BoxCars21k [17], and CompCars [23]. VeRi-776 [9] is a benchmark dateset for vehicle ReID that is collected from real-world surveillance scenarios, with over 50,000 images of 776 vehicles in total. VehicleID [8] is a surveillance dataset, which contains 26,267 vehicles and 22,1763 images in total. BoxCars21k [17] is designed for fine-grained vehicle make and model recognition. The images of BoxCars21k are ordered by identities thus can also be used for vehicle ReID. This dataset contains 21,250 vehicle identities and 63,750 images. CompCars [23] is also designed for fine-grained vehicle model classification, which consists of both web images and surveillance images. However, we only utilize its surveillance data for training. Images in this dataset are sorted by vehicle model and color annotations are also provided. We can roughly regard vehicles with specific model and specific color as a specific identity to train our ReID network.

| Dataset | #Trn ID/img | #Prb ID/img | #Gal ID/img |
|---|---|---|---|
| VeRi-776 [9] | 576/30188 | 200/1678 | 200/11579 |
| VehicleID [8] | 13164/100182 | 2400/17638 | 2400/2400 |
| BoxCars [17] | 21250/63750 | - / - | - / - |
| CompCars [23] | 1118/31148 | - / - | - / - |

Table 2. Statistics of the four datasets used in our experiment. The number of train identities and images, together with the number of query and gallery identities and images are listed.

We merge the training samples from VeRi-776 [9] and VehicleID [8], together with all the samples from BoxCars21k [17] and CompCars [23] into one large training set to train our orientation invariant network. The training set contains around 225,268 images of 36,108 identities in total. Selected samples of these datasets are shown in Fig. 7 and the statistical information are listed in Table 2.

## 4.2. Evaluation results

The proposed framework is compared with two state-of-the-art vehicle ReID approaches, *i.e*. PROVID [9] and DRDL [8], together with several conventional person ReID methods, *i.e*. Bag of Words with Color Name Descriptor (BOW-CN) [28], the LOMO feature [6], and the KEPLER method [11], which learns salient regions for constructing discriminative features. Performance evaluation is conducted on VeRi-776 [9] and VehicleID [8], and multiple evaluation metrics are applied.

For the VeRi-776 dataset, cumulative match curve (CMC) metric [3] is adopted for evaluation. For each identity, one image is random selected from all the gallery images to generate the gallery set, while the probe set remains unchanged. The random selection procedure was repeated for 100 times to obtain an average CMC result. The image-to-track metric (HIT) introduced in [9] is also evaluated, and there is no random gallery selection process for the HIT evaluation. Mean average precision (mAP) is also adopt for evaluation following [9].

For the VehicleID dataset, standard CMC metric is adopted with random gallery selection. Only the *Large* test set of VehicleID is evaluated since it is the most challenging set. During testing, one image is randomly selected from one identity to obtain a gallery set with 2,400 images, then the remaining 17,638 images are all used as probe images.

Evaluation results on both datasets are listed in Table 3. The proposed approach achieves the best performance on both datasets, which is much better than the compared methods. Three experiments are conducted to demonstrate the effectiveness of the proposed main modules. Firstly, a baseline single-branch appearance model ("Baseline") is evaluated to investigate the performance of our proposed network. Significant performance gain can be observed

| VeRi-776 | mAP | HIT@1 | CMC@1 | CMC@5 |
|---|---|---|---|---|
| BOW-CN [28] | 12.20 | 33.9 | - | - |
| LOMO [6] | 9.64 | 25.3 | - | - |
| KEPLER [11] | 33.53 | 68.7 | 48.2 | 64.3 |
| PROVID [9] | 27.77 | 61.4 | - | - |
| Baseline | 45.50 | 88.66 | 62.8 | 86.7 |
| Ours | 48.00 | 89.43 | 65.9 | 87.7 |
| Ours + ST | **51.42** | **92.35** | **68.3** | **89.7** |

| VehicleID | CMC@1 | CMC@5 |
|---|---|---|
| KEPLER [11] | 45.4 | 68.9 |
| VGG + Triplet Loss [8] | 31.9 | 50.3 |
| VGG + CCL [8] | 32.9 | 53.3 |
| Mixed Diff + CCL [8] | 38.2 | 61.6 |
| Baseline | 63.2 | 80.6 |
| Ours | **67.0** | **82.9** |

Table 3. Experiment results of the proposed method and other compared methods on the VeRi-776 dataset and VehicleID dataset. Multiple evaluation criteria are adopted, including mAP, HIT accuracy, and CMC accuracy. "Baseline" refers to our single main-branch model without region features, "Ours" denotes the proposed orientation invariant network, and "Ours+ST" indicates the overall pipeline with the proposed spatio-temporal regularization model.

compared with other methods. Secondly, the proposed orientation invariant network ("Ours") is tested by using the proposed orientation invariant appearance features. Region features are introduced and the performance can be further improved. Finally, the overall framework ("Ours+ST") is evaluated by adding the spatio-temporal regularization module. This experiment is only conducted on VeRi-776 since no spatio-temporal information is provided on VehicleID. Experimental result demonstrates that the regularization module results in significant improvements.

Note that the HIT@1 results of the proposed method in Table 3 outperform existing methods by large margins. This is because the HIT@1 metric is based on image-to-track search, and all gallery images (which contain multiple ground truth results) are searched to identify the probe identity. If there exists one gallery image of this identity that shares similar orientation as the probe image, such gallery images can be easily found. In this case, CMC should be a more proper metric for vehicle ReID evaluation, since the image-to-track search may always obtain search results with similar orientation as the probe image.

# 5. Ablation Study

## 5.1. Investigations on the Key point Regressor

In this section, the proposed key point regressor is thoroughly investigated, in terms of the regression accuracy and relationship between landmarks and orientations. In order to train and evaluate the key point regressor, locations of

| Models | $r_0 = 5$ | $r_0 = 3$ |
|---|---|---|
| L2-Loss | 90.50 | 87.4 |
| Cross-Entropy Loss | 92.05 | 88.8 |

Table 4. Evaluation results of the landmark regressor.

(a) (b)

Figure 8. (a) Annotation examples. (b) Regression results.

20 key points are annotated manually on the images of the whole VeRi-776 [9] dataset and some annotation results are shown in Fig. 8 (a). During the testing stage, response maps of the testing images are extracted and the key points are predicted as the locations with maximum response value. If the distance between the regressed landmark location and the ground truth location is smaller than a threshold $r_0$, this key point is considered as correctly predicted, otherwise the key point is wrongly predicted. Invisible key points are ignored in the evaluation step since they are expected to be handled by the proposed orientation invariant module. The prediction accuracy of visible key points are listed in Table 4, and two loss functions are adopted to train the landmark model. We can observe that $88.8\%$ key points can be correctly predicted within $r_0 = 3$ pixels to the ground truth (the final response map is of size $48 \times 48$). Some key point prediction results are shown in Fig. 8 (b).

Investigations are also conducted on the relationship between key point locations and orientation classes. Since no orientation information is provided for the VeRi-776 dataset, four orientations, *i.e.* front, back, left, right, are manually annotated for the VeRi-776 dataset. With the annotated training images, a vehicle orientation classifier is trained by using the 20 landmark response maps as input. The trained classifier yields a $93.2\%$ accuracy on the testing images, which demonstrates our key point response maps contain sufficient information to infer vehicles' orientation. It also validates that clustering landmarks by orientation is reasonable in the proposed orientation-based region proposal module.

## 5.2. Comparison with Attention Mechanism

Besides the proposed orientation-based region proposal module, attention mechanism is another possible way to select the salience regions and to obtain local region features. Experiments are conducted to compare the soft attention mechanism and the proposed orientation based region proposal framework.

We follow the standard strategy as [24] to implement the

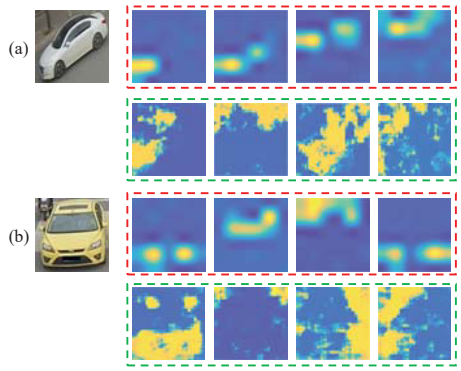


Figure 9. Comparison between the learned salience masks by the attention mechanism (in red boxes) and the orientation based region proposals (in green boxes).

attention module. $1 \times 1$ convolution layers are employed to produce salience masks from the input feature maps. For these salience masks, locations with larger value represent salience regions that are useful for feature extraction. The salience masks are then passed through a sigmoid nonlinear unit, yielding values in the range $(0, 1)$. Finally, element-wise multiplication is applied between the input feature maps and the salience mask to output the local feature maps of the salience regions.

The compared attention network is similar with the proposed pipeline in terms of network architecture. The only difference is that the 4 orientation-based region proposals are replaced with $N$ attention masks. The attention module takes the feature maps $f_0^1$ as input and output $N$ attention masks. Experiments are conducted by setting the number of attention mask $N = 2, 4, 8$ for comparison and CMC results are reported in Table 5.

The compared attention network and the proposed orientation invariant network is different. In attention network, the salience regions are learned by the attention module automatically, while in our proposed network the region masks are defined based on orientation information (four side faces of the vehicle with different landmark points) and the landmarks are regressed by the trained landmark regressor.

The orientation based region proposals and the salience masks learned by the attention modules are visualized in Fig. 9, which are also different. Attention masks are reasonable but not stable, *i.e.*, different attention blocks may provide quite similar salience masks (the first and the last mask in Fig. 9(b)). However, our orientation based region proposals is much more stable, because they are designed to focus on different faces. Experimental results in Table 5 also demonstrate that the orientation based region proposals outperform attention mechanism.

| Models | VeRi-CMC@1 | VehicleID-CMC@1 |
|---|---|---|
| attention-2branch | 63.6 | 64.3 |
| attention-4branch | 64.8 | 65.6 |
| attention-8branch | 62.7 | 63.8 |
| Ours | 65.9 | 66.6 |

Table 5. Quantitative results by replacing orientation based region proposal with attention masks.

| Models | VeRi-CMC@1 | VID-CMC@1 |
|---|---|---|
| base | 88.66 | 63.2 |
| global+KISSME [4] | 89.02 | 63.5 |
| global+MLAPG [7] | 87.89 | 63.1 |
| global+Zhang *et al*. [26] | 89.11 | 63.8 |
| final | 89.43 | 67.7 |
| final+KISSME [4] | 89.75 | 67.8 |
| final+MLAPG [7] | 88.89 | 67.1 |
| final+Zhang *et al*. [26] | 90.05 | 68.0 |

Table 6. Experimental results with and without metric learning. The first group of experiments are based on our global features while the second group are based on our global + local features

### 5.3. Metric Learning Methods

In person Re-ID methods, metric learning is usually utilized to refine the final search results. In this section, we tested some popular metric learning methods utilized in person Re-ID methods, including KISSME [4], MLAPG [7] and Zhang *et al*. [26]. We utilized the three metric learning algorithms on features extracted from both our baseline model (global feature) and final model (global + local feature) and conducted two groups of experiments. The results on both the VeRi and VehicleID (VID) datasets are listed in Table 6:

As show in the table, metric learning methods do lead to additional performance gain based on our features. Note that in the second group of experiments performance gains by metric learning methods are smaller, which suggest that our global+local features are more discriminative than the global ones.

## 6. Conclusion

In this paper, a novel framework is proposed for vehicle ReID, which consists of two main components, *i.e.* the orientation invariant feature embedding and the spatial-temporal regularization. Local region features are extracted based on the locations of key points. Th features are aligned and combined to form orientation invariant feature representations. Spatio temporal regularization is adopted for refining the retrieval results. The proposed framework is evaluated on two public vehicle ReID datasets and state-of-the-art performance is achieved. Detailed investigations of the proposed methods are conducted in terms of the landmark regressor and the comparisons with attention mechanism.

## References


[1] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 2

[2] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 2

[3] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, 2007. 6

[4] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012. 8

[5] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 2

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, June 2015. 2, 6, 7

[7] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015. 8

[8] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016. 2, 6, 7

[9] M. T. M. H. Liu X., Liu W. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, 2016. 2, 6, 7

[10] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 4, 5

[11] N. Martinel, C. Micheloni, and G. L. Foresti. Kernelized saliency-based person re-identification through multiple metric learning. *IEEE Transactions on Image Processing*, 24(12):5645–5658, 2015. 6, 7

[12] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2

[13] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 2, 3

[14] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692, 2014. 2

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2

[16] Z. Shi, T. M. Hospedales, and T. Xiang. Transferring a semantic representation for person re-identification and search. In *CVPR*, 2015. 2

[17] J. Sochor, A. Herout, and J. Havel. BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition. In *CVPR*, June 2016. 2, 6

[18] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 2

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, June 2015. 4

[20] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 2

[21] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016. 2

[22] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. *arXiv preprint arXiv:1604.07528*, 2016. 2

[23] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, June 2015. 2, 6

[24] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, June 2016. 7

[25] D. Zapletal and A. Herout. Vehicle re-identification for automatic video traffic surveillance. In *CVPR Workshops*, pages 25–31, 2016. 2

[26] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016. 8

[27] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 2

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2, 6, 7