Part-regularized Near-duplicate Vehicle Re-identification

Bing He\textsuperscript{1} Jia Li\textsuperscript{1,3,4*} Yifan Zhao\textsuperscript{1} Yonghong Tian\textsuperscript{2,3}
\textsuperscript{1}State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
\textsuperscript{2}National Engineering Laboratory for Video Technology, School of EE&CS, Peking University
\textsuperscript{3}Peng Cheng Laboratory, Shenzhen, China
\textsuperscript{4}Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China
\textsuperscript{1}\{bing, jiali, zhaoyf\}@buaa.edu.cn \textsuperscript{2}yhtian@pku.edu.cn

Abstract

Vehicle re-identification (Re-ID) has been attracting more interests in computer vision owing to its great contributions in urban surveillance and intelligent transportation. With the development of deep learning approaches, vehicle Re-ID still faces a near-duplicate challenge, which is to distinguish different instances with nearly identical appearances. Previous methods simply rely on the global visual features to handle this problem. In this paper, we proposed a simple but efficient part-regularized discriminative feature preserving method which enhances the perceptive ability of subtle discrepancies. We further develop a novel framework to integrate part constraints with the global Re-ID modules by introducing an detection branch. Our framework is trained end-to-end with combined local and global constraints. Specially, without the part-regularized local constrains in inference step, our Re-ID network outperforms the state-of-the-art method by a large margin on large benchmark datasets VehicleID and VeRi-776.

1. Introduction

Given a query image of a vehicle identity, vehicle re-identification task aims to retrieve all the images of this identity from a large image database which typically captured from a large camera network. With the proposals of large dataset \cite{14, 12, 27} and the development of deep learning algorithms \cite{24, 36}, recent models have gain remarkable success in the past decade. The re-identification of vehicles has a great potential to contribute to the urban security surveillance and intelligent transportation.

Considering the inconspicuous divergences among different instances, vehicle re-identification is still a very challenging task, especially with the large amount of dataset. To address this Re-ID task, many deep learning models \cite{27, 1} relied on global information have been proposed in the past decades. One intuitive solution is to reduce the distances of identical vehicle images and enlarge the distance of different ones with learning approaches. To better measure the distance, previous works \cite{12} mainly use deep metric learning to directly embed the raw image into an Euclidean space where the distance can be directly used as similarity scores between two vehicles. Weinberger et al. \cite{25} explore the topic of metric learning to perform k-nearest neighbor classification and propose the Large Margin Nearest Neighbor loss (LMNN). FaceNet \cite{20} improved the LMNN loss into a modified triplet loss which directly optimize the final distance metric and can be applied in re-identification and face recognition tasks. Although these works reach remarkable success in vehicle re-identification tasks, they usually get confused when these vehicle have inconspicuous differences. \textit{e.g.}, see Fig. 1 (a).
To handle this problem, recent works resort to additional license plate and spatial-temporal information. Liu et al. [14] introduce license plate recognition into Re-ID task. The license plate recognition usually fails in unconstrained environment due to the various viewpoints and changeable illuminations. However, owing to the privacy and security considerations in vehicle re-identification task, the plate information is inaccessible in the public benchmarks. Besides, some other methods [21, 24] rely on extra spatial-temporal information to explore the final retrieval results.

In this paper, we explore the near-duplicate phenomenon in vehicle re-identification. As illustrated in Fig. 1 a), different vehicles usually share similar geometric shapes and appearances which can be hard to distinguish by deep models. While the details from these near-duplicate vehicles have arresting variances in local features such as brands and tags in windows which are easily recognized by human beings, see Fig. 1 b). To handle the near duplicated phenomenon in vehicle re-identification task, we propose a part-regularized approach which integrates the local and no-local features into a unified architecture. To avoid the vanish of local features, we enhance the perception of local information of regularized parts in deep learning networks. Inspired by ROI (region of interest) in object detection, we adopt ROI receptive module to capture the local information. We develop a simple but effective ROI projection approach to combine detection branch with our Re-ID task. After combining these features, we further developed a local and no-local classification loss. To summarize, the contribution of our work is three-fold:

- We design an effective representation learning framework by jointly considering local and global representations.
- We propose a part-regularized approach to enhance the discriminative capability of global features for vehicle re-identification.
- We conduct extensive experiments to show that the proposed approach outperforms state-of-the-art: VehicleID [12] by 57% in rank-1, 23% in rank-5, VeRi-776 [14] by 48% in mAP, 2.1% in HIT@1 and 9.6% in HIT@5.

The rest of this paper is organized as follows: Sec. 2 reviews the related works, Sec. 3 gives the problem statement of vehicle re-identification and explains the details of our part-regularized model. Qualitative and quantitative experiments are presented in Sec. 4 and we finally conclude our paper in Sec. 5.

2. Related Work

Vehicle Re-ID. The vehicle re-identification task has gained more and more attention in recent years. Li et al. [12] proposed a benchmark dataset VehicleID and a pipeline which use Deep Relative Distance Learning (DRDL) to project vehicle images into an Euclidean space, where the distance can directly measure the similarity of two vehicle images. Liu et al. [14] proposed another dataset, which called VeRi-776, and build a coarse-to-fine progressive search framework through utilizing the visual appearance, license plate and spatial-temporal information. VeRi-776 contains rich annotations including vehicle types, colors, brands, license plate and spatio-temporal information. Wang et al. [24] explored vehicle viewpoint attribute and proposed orientation invariant feature embedding module. The orientations information are extracted by 20 vehicle key points locations. Shen et al. [21] pushed spatial-temporal idea further and proposed Visual-spatial-temporal Path Proposals method. Yan et al. [27] model the relationships of vehicle images as multi-grain list and proposes two ranking methods, generalized pairwise ranking and multi-grain based list ranking to address this problem, and contributed two high-quality and well-annotated vehicle datasets VD1 and VD2, which are collected from two different cities with diverse annotated attributes. While Lou et al. [15] resort to adversarial learning to generate cross views of new examples.

Person Re-ID. Person re-identification aims to retrieve all the images of the query individual from a large scale image database. The person re-id methods can be roughly categorized into two groups, classification methods and Siamese methods based on triplet comparisons. Li et al. [7] proposed a multi-scale context aware network that can capture knowledge of the local context. Xiao et al. [26] proposed a model to learn deep feature representations from multiple dataset with Convolutional Neural Networks. Their experiment shows that some neurons learn representations shared across all datasets, while some others are effective only for a specific domain. Su et al. [22] proposed a pose-driven convolutional neural network to address the large pose deformations and the complex view variations problem. AlignedReID [31] learns a global feature but performs part alignment during training. Local feature is extracted by horizontal pooling from each row, without requiring additional supervision or pose estimation.

Discriminative part localization. Discriminative part localization has been studied for a long time by many community such as fine-grained recognition [5, 10, 18, 29, 30], face recognition [37, 16, 17, 33, 23] and person re-identification [26]. After deep learning dominate computer vision community, hand-craft part features for fine-grained recognition has been drooped. Many works [34, 8] in person re-identification exploited human body parts to learn robust representations. Li et al. [8] proposed to learn and localize deformable pedestrian parts using Spatial Transformer Networks(STN). Using semantic segmentation’s a-

3. Methodology

3.1. Problem Statement

Given a query image, the target of vehicle re-identification is to compute the similarity score between this query image and all the other images in the gallery. Define the training set as \( \{ x_i, y_i \}_{i=1}^{N} \). Each vehicle image \( x_i \) is labeled with identification label \( y_i \) with the total number of \( N \) training images. The training Images and identification labels are denoted as \( x \) and \( y \) respectively. The desired similarity between probe \( p \) and gallery image \( g \) is defined as \( M(\phi(p; \theta), \phi(g; \theta)) \), where \( \phi(\cdot; \theta) \) is the feature extraction function which usually denotes a common deep encoder, and \( M(\cdot) \) is a metric defined in the feature space. The most important question is how to learn the feature extraction function \( \phi(\cdot; \theta) \). Previous works use classification method to learn parameters \( \theta \) in function \( \phi(\cdot; \theta) \), from which the optimization target can be defined as

\[
\arg \min_{\theta} \mathbb{E} \left( \phi(x; \theta)^\top w, y \right),
\]

where \( \phi(x; \theta) \) is the feature extracted by deep neural network with parameter \( \theta \), \( w \) is the parameter to project the features into predicted labels. \( \mathbb{E}(\cdot) \) is the cross entropy loss. As discussed before, the equations above only optimize the global feature and become easy to ignore subtle visual cues. To handle this problem, We introduce part information and propose a novel local feature based optimization target which is defined as

\[
\arg \min_{\theta} \mathbb{E} \left( \phi(x; \theta)^\top w_g, y \right) + \sum_{p \in P} \lambda_p \mathbb{E} \left( (\phi(x; \theta) \circ M_p)^\top w_l, y_p \right),
\]

where \( w_g \) is the parameter to project the global feature into predicted identification label. \( w_l \) is the local parameter that project the local part feature into predicted part label. \( M_p \) is the part location that can be used to extract local feature from global feature. \( \circ \) is the local feature extraction operation. This formulation introduce the part constrain to the re-id task and force the network preserve the local part cue to recognize parts. Details will be explained in section 3.2.
There are still some unsolved problems in Eq. (2). First, the part set \( \mathcal{P} \) is not defined which means we don’t know which part should be used. Second, the part location \( M_p \) need to be extracted. Third, \( y_p \), which is the part label, should be determined. In the next subsection, we will explain our network structure to address these problems.

### 3.2. Part-Regularized Re-ID

In this section, we introduce part regularized (PR) constraints into the vehicle re-identification task. Our framework consists of two components, a global module to conduct Re-ID categorization and a local part-regularized module to encourage correct classification of the identified parts. To preserve better context information, which is very crucial for the near-duplicate problem, we adopt bounding box detection network for part localization. We will explain the details of the two main components in this section and describe training scheme in section 3.3.

**Part definition.** We select three vehicle parts for our part detection module, lights, including front light and back light, window, including front window and back window, and vehicle brand. The vehicle head area is crucial to distinguish different vehicle models. We use the front lights to infer the vehicle head area including the brand. Different model may have extremely difﬁdence lights, we define bounding box of the light as tight bounding box contains the light but extend it to the bottom of the vehicle. This deﬁnition can preserve more context information which we ﬁnd more stable in experiment. The definition of the three parts in our model is shown in Fig. 3. We draw \( N \) local branches in Fig. 2 since our framework is ﬂexible to various deﬁnitions of vehicle parts, and we only test \( N = 3 \) parts (window, light, brand) to validate the effectiveness of this framework.

**Part detection.** To solve the second problem, we need to ﬁnd the parts location of the training images. There are many off the shelf object part localization algorithms, which can mainly categorize into two classes, detection and segmentation. Segmentation method need pixel level annotation which is difficult to get. In this paper we use a detection branch to detect the predefined vehicle parts. As shown in the Fig. 2, raw vehicle image is fed into the LocalNet (use YOLO in experiments), which has 24 convolutional layers, to get raw part detection results. A desired result is that every image get three bounding box for window, left light and right light respectively. During the training process, we ﬁnd that in some rare cases the vehicle part detection model may fail due to occlusion. To handle these invisible parts in a speciﬁc vehicle image, we refer to the rest images of the same vehicle and compute the average locations of the missing parts. After that, these average part locations are used as the pseudo detection results of this speciﬁc image to facilitate the subsequent training process.

![Figure 3. The part definition of our model. The first row shows the vehicle window part in both front and back view. Vehicle lights are shown in the second row. We extend the bounding box of the lights to the bottom of the vehicle to preserve more context information. The head and rear area of the vehicle containing the vehicle brand is defined as vehicle brand part.](image-url)
modified as
\[
\arg \min_{\theta} \mathbb{E} \left( \phi(x; \theta)^\top w_g, y \right) + \sum_{p \in P} \lambda_p \mathbb{E} \left( \left( \phi(x; \theta) \odot M_p \right)^\top w_l, y \right),
\]
where \( y \approx y_p \), now we can use vehicle identification label to optimize our model.

### 3.3. Training Scheme

Both of our part localization module and part feature extraction and aggregation module can be trained end to end using backpropagation. We adopt the successful YOLO network [19] as our backbone of LocalNet. In training steps, first we train the part detection module and extract all the part locations of the training images. Part information of the test images was not extracted since we don’t use local feature branch at test stage. For VehicleID and VeRi-776, we adopt the transfer learning scheme and use the ImageNet pretrained weights for backbone network GlobalNet(ResNet-50). Then we use the optimization function defined in Eq. (3) with a initial learning rate \( lr = 0.01 \) with exponential learning rate schedule to fine-tune the whole feature extraction module, including global and local branch.

### 4. Experiment

#### 4.1. Datasets and Evaluation Metric

We evaluate our proposed model on two public large-scale vehicle re-identification datasets, VehicleID and VeRi-776.

VeRi-776 is a benchmark dataset for vehicle re-id task. It contains about 50,000 images of 776 vehicles labeled with rich attributes, e.g. types, colors, brands, license plate annotation and spatiotemporal relation annotation. Each vehicle was captured by various cameras with different view points. The short coming of this dataset is that the number of identities is relatively small, in test stage it is very easy to distinguish each vehicle just based on model information. We use the official dataset settings and adopt mAP, HIT-1 and HIT-5 to evaluate our proposed model.

VehicleID is another benchmark with larger data volume. VehicleID is captured by multiple non-overlapping cameras and there are 221,763 images of 26,267 vehicles in total. Each image is either captured from the front view or back view. In VehicleID, only 250 vehicle models are included, which means many different identities share same vehicle model, near-duplicate problem appears. We use mAP to evaluate our method on three subset(i.e. small, medium and large) of the testset.

There are no bounding box annotations of the vehicle parts in both the VehicleID and VeRi-776 dataset. Therefore, we randomly select 500 vehicle images from the VehicleID dataset and label three vehicle parts with bounding boxes (window, light and brand), and these images are used to train the YOLO model. The trained model shows impressive detection results on both VehicleID and VeRi-776 dataset, implying a good generalization ability. The annotation process is also quite efficient and costs only 4 hours of one person in annotating all the 500 images.

The mean average precision (mAP) and cumulative match curve (CMC) are adopted in our experiments. For VeRi-776, the image-to-track metric HIT@1 and HIT@5 is also reported. The CMC curve shows the probability that the image of the probe identity appears in different-sized retrieved list. CMC can be calculated as
\[
CMC@k = \frac{\sum_{i=1}^{N} m(q_i, k)}{N},
\]
where \( N \) is the number of queries and \( m(q_i, k) \) equals to 1 if \( q_i \) appears in the top-k of the rank list. The number of ground truth image of a probe should be exactly 1 in order to use the cumulative match curve. The precision measures the accurate of the prediction, the average precision for each query q can be calculated as
\[
AP(q) = \sum_{k=1}^{N} P(k) \Delta r(k),
\]
where $P(k)$ is the precision at a cutoff of $k$ images, $N$ is the total number of images in the gallery, and $\Delta r(k)$ is the change in recall that happened between cutoff $k - 1$ and cutoff $k$. The mean average precision for all query images is determined by

$$mAP = \frac{\sum_{q=1}^{N} AP(q)}{Q},$$

where $Q$ is the total number of queries.

4.2. Experiment Setup

We use ResNet-50 as the backbone network for feature extraction. We apply average global pooling[11] on the global feature map followed with a $1 \times 1$ convolutional layer to extract the final 256-d global feature vector. Euclidean distance (L2) was adopted to compute similarity score between query and gallery images at both training and testing stage. It is worth mentioning that we only use global branch at test stage because during the experiment we found that fusing global and local part features yields similar performance compared to just using global branch. Which mean our model does not need part detection at test stage.

4.3. Comparison with State-of-the-art

The proposed method is compared with state-of-the-art vehicle re-identification methods on two datasets.

**VehicleID.** For VehicleID dataset, testing data is split into three subsets ordered by their size. For each test dataset split, one image of each vehicle identity is selected and put into the gallery set. The rest images are all probe queries. In this setting each vehicle identity has many query images but have only one gallery images, so the cumulative match curve (CMC) metric is adopted for evaluation. Table 1 and Table 2 show performance comparisons on VehicleID. Our model outperform all the existing method. OIFE [24] and VAMI [36] exploit the vehicle view information use the view invariant feature to roughly align the vehicle image. Those view align methods are useful when distinguish different vehicles from difference vehicle model, but they can’t address the near-duplicate problem since appearance of same viewpoint of the near-duplicate vehicles are still fairly similar. It need more detail cues than vehicle view informations to distinguish the near-duplicate vehicle.

**VeRi-776.** The cross-camera search is performed followed the official settings in [14]. At test stage, each image of a vehicle from every camera is selected as probe image and used to search for tracks of the same vehicle in other cameras. That means evaluation for VeRi-776 is conducted in an image-to-track fashion, in which the probe is an image, while the targets are images in the track. The problem is how to define the similarity between a query image and a gallery track. Following the settings in [14], the similarity is defined as maximum similarity between a query image and all images in the track. The image-to-track evaluation results is shown in Table 3. Fact+Plate+STR [14], Siamese+Path [21] and OIFE+ST [24] relies on the spatiotemporal information in Veri-776 Dataset. Fact+Plate+STR [14] uses the additional license plate informations. Other methods only rely on the visual information including ours. Our Part-regularized model outperform all the existing method on mAP metric.

![Figure 5. Class activation map (CAM) generated by identification classification model. CAMs with (part-regularized) PR method are show in the first and third rows while NoPR are in the second and forth rows. Activation maps with PR can easily distinguish different cars by the accurate part information for near-duplicated vehicles while NoPR models are usually get confused. It is worth mentioning that the activation map without RP can attend to vehicle light or brand parts originally.](image)

| Table 1. Result of CMC@1 in VehicleID Dataset. |
|----------------------|--------|--------|--------|
| Method               | Small  | Medium | Large  |
| VGG+Triplet Loss [2] | 0.404  | 0.354  | 0.319  |
| VGG+CCL [12]         | 0.436  | 0.370  | 0.329  |
| Mixed Diff+CCL [12]  | 0.490  | 0.428  | 0.382  |
| OIFE [24]            | -      | -      | 0.670  |
| VAMI [36]            | 0.631  | 0.529  | 0.473  |
| Ours                 | 0.784  | 0.750  | 0.742  |
Figure 6. Rank list visualization. The first image with green border in each row is the query image, the rest images are retrieved from the gallery and sorted by similarity score (L2 distance). The ground-truth is marked with red border. (a) rank list result of our full model with all three part branch. The first image in each row is the query, and the rest ten images are the top ten retrieval results. (b) rank list result after removing the window branch. (c) rank list result after removing the lights and brand branches.

<table>
<thead>
<tr>
<th>Method</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG+Triplet Loss [2]</td>
<td>0.617</td>
<td>0.546</td>
<td>0.503</td>
</tr>
<tr>
<td>VGG+CCL [12]</td>
<td>0.642</td>
<td>0.571</td>
<td>0.533</td>
</tr>
<tr>
<td>Mixed Diff+CCL [12]</td>
<td>0.735</td>
<td>0.668</td>
<td>0.616</td>
</tr>
<tr>
<td>OIFE [24]</td>
<td>-</td>
<td>-</td>
<td>0.829</td>
</tr>
<tr>
<td>VAMI [36]</td>
<td>0.833</td>
<td>0.751</td>
<td>0.703</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>0.923</strong></td>
<td><strong>0.883</strong></td>
<td><strong>0.864</strong></td>
</tr>
</tbody>
</table>

Table 2. Result of CMC@5 in VehicleID Dataset.

4.4. Ablation Study

We conduct ablation study on VehicleID dataset to investigate the effectiveness of each part branch in our model. There are three local part branch in our framework, window branch, light branch and brand branch. We remove one branch at a time and retrain the whole network to evaluate the performance. Rank list visualization is also performed as shown in Fig. 6.

including those who use extra none-visual cues.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>HIT@1</th>
<th>HIT@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW-CN [35]</td>
<td>0.122</td>
<td>0.339</td>
<td>0.537</td>
</tr>
<tr>
<td>LOMO [9]</td>
<td>0.096</td>
<td>0.253</td>
<td>0.465</td>
</tr>
<tr>
<td>GoogLeNet [28]</td>
<td>0.170</td>
<td>0.498</td>
<td>0.712</td>
</tr>
<tr>
<td>FACT [13]</td>
<td>0.185</td>
<td>0.510</td>
<td>0.735</td>
</tr>
<tr>
<td>Plate-SNN [14]</td>
<td>0.157</td>
<td>0.363</td>
<td>0.466</td>
</tr>
<tr>
<td>FACT+Plate-REC [14]</td>
<td>0.186</td>
<td>0.512</td>
<td>0.736</td>
</tr>
<tr>
<td>FACT+Plate-SNN [14]</td>
<td>0.259</td>
<td>0.611</td>
<td>0.774</td>
</tr>
<tr>
<td>FACT+Plate+STR [14]</td>
<td>0.278</td>
<td>0.614</td>
<td>0.788</td>
</tr>
<tr>
<td>Siamese+Path [21]</td>
<td>0.583</td>
<td>0.835</td>
<td>0.900</td>
</tr>
<tr>
<td>OIFE [24]</td>
<td>0.480</td>
<td>0.894</td>
<td>-</td>
</tr>
<tr>
<td>OIFE+ST [24]</td>
<td>0.514</td>
<td>0.924</td>
<td>-</td>
</tr>
<tr>
<td>VAMI [36]</td>
<td>0.501</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>0.743</strong></td>
<td><strong>0.943</strong></td>
<td><strong>0.987</strong></td>
</tr>
</tbody>
</table>

Table 3. Results of mAP and HIT@1 HIT@5 in VeRi-776 Dataset.
Table 4. Results of Match Rate of ablation experiment.

<table>
<thead>
<tr>
<th>Method</th>
<th>CMC@1</th>
<th>CMC@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global+Light+Brand+Window</td>
<td>0.742</td>
<td>0.864</td>
</tr>
<tr>
<td>Global+Light+Brand</td>
<td>0.675</td>
<td>0.830</td>
</tr>
<tr>
<td>Global+Light+Window</td>
<td>0.710</td>
<td>0.887</td>
</tr>
<tr>
<td>Global+Window+Brand</td>
<td>0.726</td>
<td>0.851</td>
</tr>
<tr>
<td>Global+Window</td>
<td>0.707</td>
<td>0.832</td>
</tr>
<tr>
<td>Window+Light+Brand</td>
<td>0.687</td>
<td>0.829</td>
</tr>
<tr>
<td>Baseline (w/o parts)</td>
<td>0.645</td>
<td>0.800</td>
</tr>
</tbody>
</table>

Table 5. Influences of different resolutions in VehicleID dataset.

<table>
<thead>
<tr>
<th>VehicleSet</th>
<th>Input size</th>
<th>CMC@1</th>
<th>CMC@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>128 × 128</td>
<td>0.726</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>256 × 256</td>
<td>0.784</td>
<td>0.923</td>
</tr>
<tr>
<td>Medium</td>
<td>128 × 128</td>
<td>0.685</td>
<td>0.838</td>
</tr>
<tr>
<td></td>
<td>256 × 256</td>
<td>0.750</td>
<td>0.883</td>
</tr>
<tr>
<td>Large</td>
<td>128 × 128</td>
<td>0.661</td>
<td>0.819</td>
</tr>
<tr>
<td></td>
<td>256 × 256</td>
<td>0.742</td>
<td>0.864</td>
</tr>
</tbody>
</table>

Vehicle window. As shown in Table 4, cutting off the vehicle window branch depress the re-id performance by 7 percent. Vehicle window contains the personality feature which is crucial to distinguish difference vehicle identities from same vehicle model. The visualization result confirms this point. As shown in Fig 6, almost all of the top 10 retrieval results are come from the same vehicle model. In this scenario visual cues from vehicle window become extremely important since others vehicle parts are almost the same.

Vehicle brand and light. Removing the vehicle brand or vehicle light branch also depress the performance of our model. Compared to cutting off the vehicle window branch, removing the vehicle light and brand only yields a smaller performance drop. This is because the global feature can learn some of the vehicle light and brand information originally as discussed before in Fig 5. The performance drop shows that putting explicit constrains to the neural network makes the learning process more efficient.

Global branch. We cutting off the global branch and only use three part branch to train the network. During testing three part feature vector is extracted and fusing together to compute the similarity score. The performance drops a lot unsurprisingly. The other part like vehicle body and wheels are useful when distinguish two vehicle identities. The global branch is response to extract those discriminative information.

Influences of resolution. We conduct experiments on different resolutions of input size, as shown in Tab. 5 and 6. For VehicleID dataset, we conduct experiments on three testset with different image resolutions. One intuitive observation is that images with higher resolution performs better but with a higher computation cost. Interestingly, we find that images with size 128 × 128 exhibit large performance drop especially for CMC@1 indicator, while for CMC@5 and HIT@5, images with low resolution yield feasible results.

5. Conclusions

In this paper, we explore the near-duplicate challenge which causes one of the most remarkable confusions in vehicle re-identification tasks. To enlarge the divergences between nearly identical instances, we proposed a simple but efficient part-regularized approach which enhances the local features in the original Re-ID task. Our model introduces part level constrains to the typical Re-ID framework to enhance the perceptive of subtle discrepancies, which is crucial for the near-duplicate vehicle Re-ID, not being ignored during forward propagation and the detection ROIs on feature maps is the best practice to facilitate the local visual cues. We also conduct qualities and quantities experiments to demonstrate the effectiveness of each branch in our framework.

Acknowledgments

This work was supported in part by the National Basic Research Program of China under Grant 2015CB351806, in part by the National Natural Science Foundation of China under Grant 61672072, Grant 61532003, and Grant 61825101, and in part by Beijing Nova Program under Grant Z181100006218063.

References


