Partition and Reunion: A Two-Branch Neural Network for Vehicle Re-identification

Hao Chen[1], Benoit Lagadec[2], and Francois Bremond[1]

[1]INRIA Sophia Antipolis, 2004 Route des Lucioles, 06902, Valbonne, France
{hao.chen,francois.bremond}@inria.fr
[2]European Systems Integration, 362 Avenue du Campon, 06110, Le Cannet, France
benoit.lagadec@esifrance.net

Abstract

*The smart city vision raises the prospect that cities will become more intelligent in various fields, such as more sustainable environment and a better quality of life for residents. As a key component of smart cities, intelligent transportation system highlights the importance of vehicle re-identification (Re-ID). However, as compared to the rapid progress on person Re-ID, vehicle Re-ID advances at a relatively slow pace. Some previous state-of-the-art approaches strongly rely on extra annotation, like attributes (*e.g.*, vehicle color and type) and key-points (*e.g.*, wheels and lamps). Recent work on person Re-ID shows that extracting more local features can achieve a better performance without considering extra annotation. In this paper, we propose an end-to-end trainable two-branch Partition and Reunion Network (PRN) for the challenging vehicle Re-ID task. Utilizing only identity labels, our proposed method outperforms existing state-of-the-art methods on four vehicle Re-ID benchmark datasets, including VeRi-776, VehicleID, VRIC and CityFlow-ReID by a large margin.*

1. Introduction

Vehicle re-identification targets at identifying the same vehicle across a non-overlapping camera network. It has diverse real-world applications. First, vehicle Re-ID can assist the police in the fight against crime. Moreover, it can also help city planners to get a better understanding of traffic flow. Vehicle Re-ID attracts increasing attention in computer vision community.

License plate number is naturally the most crucial information for people to recognize a vehicle, which has been widely applied in real-world traffic flow analysis systems. However, due to low resolution cameras, occlusions and non-optimal viewpoints, license plate number recognition performance drops dramatically in real-world scenarios. Thus, a visual appearance based vehicle Re-ID can increase the performance of a traffic flow analysis system in cases where license plate number is not observable.

Figure 1. Examples of similar vehicles with their identity numbers in VeRi-776 dataset [20]. These pickup trunks have almost identical model and color, which makes vehicle Re-ID extremely challenging. One possible solution is to extract more discriminative local features from distinct regions.

Since both person and vehicle Re-ID belong conceptually to image retrieval problem, some common used strategies in person Re-ID can also be useful in vehicle Re-ID. To understand what can be common strategies in Re-ID tasks, we need to analyze similarities and differences between these two types of Re-ID tasks. Vehicle Re-ID and person Re-ID face several similar challenges:

- High appearance variance resulting from different illumination levels, viewpoint changes and diverse camera properties.

- Occlusions by other people/cars or static objects (trees, road sign panels, *etc*.)
- Bounding box misalignment from imperfect automated detection systems.

What makes vehicle Re-ID different from person Re-ID:

- Given that there are a limited number of vehicle colors and types, the low diversity in a vehicle Re-ID dataset make it more challenging. Figure 1 shows some examples. Therefore, salient local information is more important in vehicle Re-ID task.
- Human Body is vertically symmetrical and can be partitioned into head, torso, legs and feet along the height dimension, which makes height-wise partition more useful. Front and rear view of a car are also vertically symmetrical. But the side view is asymmetrical and can be roughly divided into hood, doors, trunk, *etc*. along the horizontal axis. On a multi-view vehicle Re-ID dataset, both height-wise and width-wise partition are helpful.

The common approach in Re-ID tasks is to build an appearance signature for each candidate image, on which we can measure similarities between query and gallery images. People usually leverage both global features on whole body and local features on partitioned body parts to tackle misalignment and occlusion problems in person Re-ID tasks. Since salient local information is more important in vehicle Re-ID, we adopt multiple partitions along 3 dimensions (height, width and channel) in feature maps to insure that our network can learn more salient local information. These partitions are illustrated in Figure 2. Meanwhile, both height and width belong to spatial dimensions, and as a consequence features extracted from the same location on the feature map can be considered twice in two spatial partitions. To avoid this issue, we split last layers of a backbone network into 2 branches, from which 2 feature maps are generated. Then, height-wise and width-wise partition are implemented separately on these 2 feature maps.

In summary, our major contribution is threefold:

1. We propose a novel 3-dimension partition strategy to extract more local features from each dimension of images.
2. By leveraging a 2-branch structure, we split our network into one "Height-Channel" branch and one "Width-Channel" branch, which avoids certain spatial features being considered twice.
3. For the challenging vehicle Re-ID, we propose an end-to-end trainable two-branch Partition and Reunion Network (PRN). In PRN, global and local features are combined together to build more robust visual signatures.

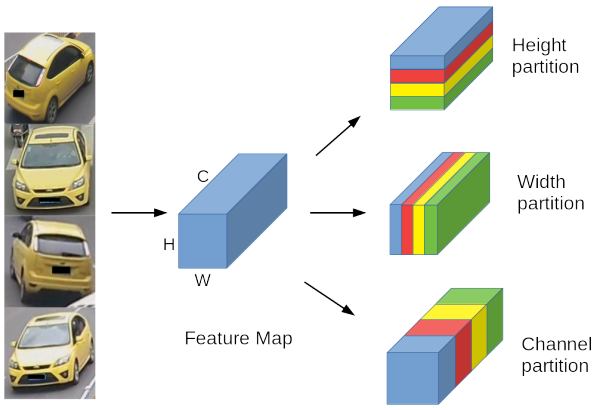

Figure 2. Illustration of our proposed 3-dimension partition strategy. H, W, C refer respectively to Height, Width and Channel dimensions of a feature map.

Results of experiments conducted on 4 large scale vehicle Re-ID datasets show that our proposed model PRN significantly surpasses state-of-the-art methods. We further conduct ablation studies to separately verify the effectiveness of each component in PRN.

2. Related Work

We firstly review some related studies conducted for person Re-ID. Since person Re-ID and vehicle Re-ID are similar tasks, techniques from one task can usually be adapted to the other one. Then, we mainly compare different approaches on vehicle Re-ID.

Person Re-ID: In computer vision community, person Re-ID has always been a hot topic and follows a common pipeline: learning from training data to build appearance signatures for query and gallery images, then comparing quantitatively the similarity between each other. In [18], authors adopt a pair-wise Siamese network architecture to treat Re-ID as a verification task. Instead of using pair-wise matching, triplet loss considers three samples simultaneously by pushing away negative sample from anchor, which has proven to be more efficient for identification tasks. Cheng *et al*. [4] adopt triplet loss to a part-based model. Hermans *et al*. [10] introduce variants of the classic triplet loss, *i.e*., Batch Hard and Batch All triplet losses. These variants confirm the effectiveness of the triplet loss for person Re-ID tasks. Our proposed model PRN combines Batch Hard triplet loss with softmax cross-entropy loss with the objective of boosting PRN's performance on both hard and normal samples.

Another popular strategy for obtaining a better performance in person Re-ID task is to train network on distinct partitioned parts and to combine local features with global features. Both hand-crafted feature based models [8, 6, 16] and deep learning based models [30, 25, 7] have taken it into consideration. To tackle the high similarity problem in Figure 1, we leverage multiple partitions jointly at 3 dimensions on feature map to fully exploit local features.

Vehicle Re-ID: As compared to person Re-ID, vehicle Re-ID is a relatively understudied topic. Prior to deep learning approaches, people usually work on hand-crafted features. In [22], authors propose to recognize vehicles with their license plate numbers and to infer their trajectories for a better understanding in urban traffic flow. Zapletal *et al*. [31] propose a real-time Re-ID model using color histograms and histograms of oriented gradients. However, hand-crafted feature based approaches usually lose performance on large scale datasets. To bridge the gap between these two kinds of features, Liu *et al*. [20] propose the fusion of multiple features *e.g*., color, texture, and deep learned semantic features.

With the rapid development of deep learning techniques in computer vision community, neural network based models have become a mainstream for large-scale vehicle Re-ID. In [21], Liu *et al*. propose a large-scale benchmark VeRi-776 and enhance the performance of their previous model FACT in [20] with a Siamese network for license plate recognition and a spatio-temporal property based re-ranking. Shen *et al*. [24] propose a two-stage framework that incorporates a Siamese-CNN based network for matching visual appearance and a LSTM based path inference mechanism. Both of them strongly depend on spatio-temporal information available on VeRi-776 dataset, which makes it impossible to be implemented on datasets without these information, *e.g*., VehicleID [19]. In [35], authors leverage viewpoint information to build a viewpoint sensitive framework and transform the single-view features into a global multi-view feature representation. The problem is that more extra annotation is needed. Recent state-of-the-art appraoches are [2, 17]. Using solely identity labels, they both adopt triplet loss in their framework for learning better visual embeddings. In [2], authors conduct a group-based clustering on a VGGM network [3], while in [17] authors adopt MobileNet [11] to reduce time complexity. To extract more discriminative local features and finally build robust view-point invariant signatures, we employ a deeper backbone network (ResNet-50) and use local features extracted from multiple partitions to compensate for the absence of extra annotation.

3. Proposed Method

In this section, we first present the general architecture of our proposed PRN. Next, more details about key components in PRN are discussed in the following subsections. At the end, we focus on the loss functions that are used to train our model.

3.1. General architecture of PRN

Figure 3 shows the general architecture of PRN. Input images are sampled into batches. Each batch contains n_{id} different identities, where we randomly select n_{img} samples from images of same identity. Thus, the batch size equals $n_{id} * n_{img}$. We duplicate last layers of backbone network and split it into 2 branches in order to increase the independence of learned spatial features. Here we name our 2 branches Height-Channel Branch and Width-Channel Branch. When an image is fed into backbone network, 2 feature maps are generated from the 2 branches. In Height-Channel Branch, the feature map is partitioned into 4 horizontal strips and 4 channel groups. 4 is a relatively moderate number of partitions for a vehicle image. Too few partition parts make local features close to global features, while too many partition parts reduce global feature weight in the final appearance signature. Similarly, in Width-Channel Branch, the feature map is partitioned into 4 vertical strips and 4 channel groups. Next, a Global Max Pooling (GMP) operation is conducted on each partitioned map as well as on entire feature maps. Since dimensions of feature vectors after GMP are different, we use a 1*1 convolution layer on each vector to unify dimensions of the vectors to 256. Each 1*1 convolution layer is followed by a batch normalization (BN) layer [14]. Fully connected layers (FC) are then employed as classifiers. Outputs of FC layers are fed into 18 softmax losses. 2 global feature vectors are fed into 2 triplet losses. Parameters in the network are updated by a combination of the softmax losses and the triplet losses. All the feature vectors (Dim=256) after BN layers are concatenated together as an appearance signature (Dim=256*18) for inference .

3.2. Key components in PRN

Backbone Network. Diverse backbone networks have been adopted in previous vehicle Re-ID models, such as VGG_CNN_M_1024 [3], MobileNet [11], ResNet [9], *etc*. Conceptually, any CNN neural network designed for image classification can be adjusted as our backbone network. In our proposed PRN, a ResNet-50 is used for its well designed architecture and competitive performance. Following modification strategies in state-of-the-art person Re-ID models [25, 27], we duplicate convolutional layers after conv4_1 layer in order to split the ResNet-50 into 2 branches. To keep more deep features, the stride of down-sampling operations in the last convolutional layer conv_5 is set to 1.

3D Partitions. Height-wise partition is a common strategy in person Re-ID models, because human body can be divided into several meaningful parts, like head, thorax, legs

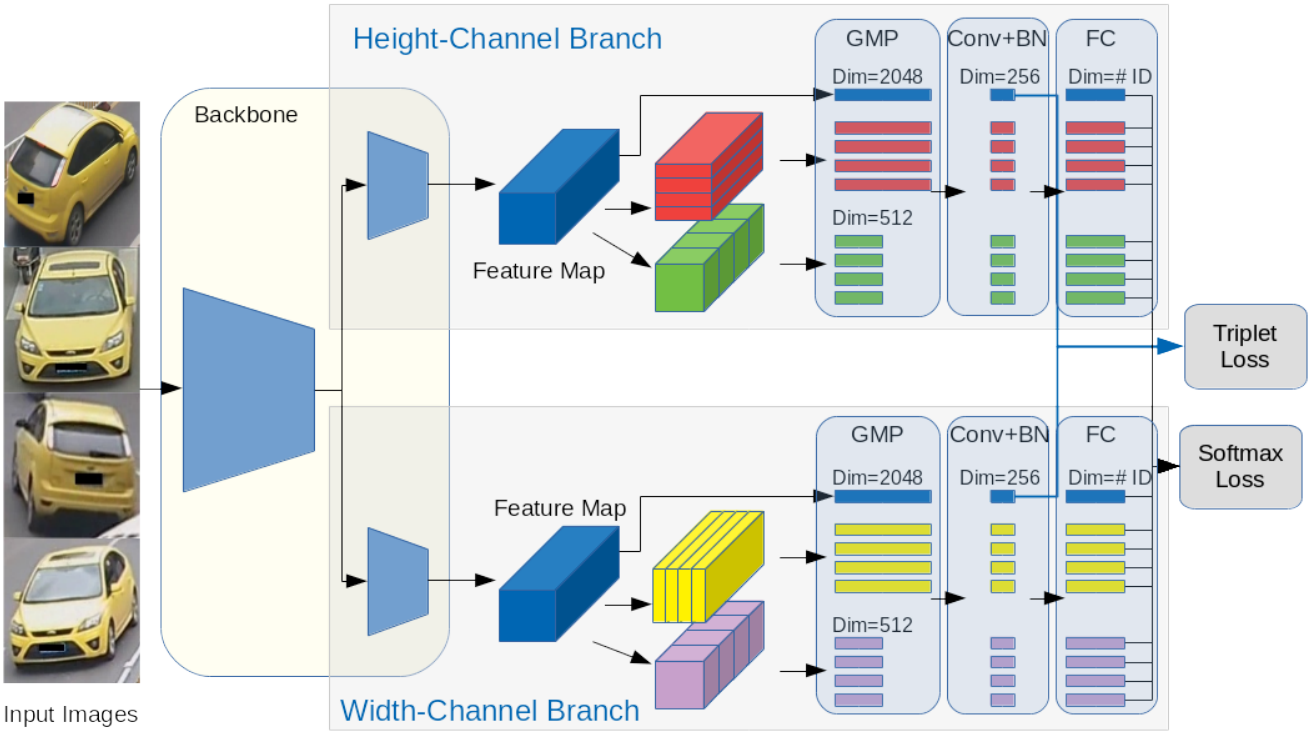

Figure 3. General architecture of our proposed model. In this paper, a ResNet-50 is used as our backbone network. Layers after conv4_1 in Resnet-50 are duplicated to split our network into 2 independent branches. GMP refers to Global Max Pooling. Conv refers to 1*1 convolutional layer, which aims to unify dimensions of global and local feature vectors. FC refers to fully connected layer. BN refers to Batch Normalization layer. In the test phase, all the feature vectors (Dim=256) after Batch Normalization layer are concatenated together as an appearance signature (Dim=256*18).

and feet. When horizontal flip operation is applied, it makes no sense to consider width partition. Unlike the human body, a car body can be roughly divided into ceiling, windshield, header panel, wheels, *etc*. on the vertical axis, and into hood, doors, trunk, *etc*. on the horizontal axis. Thus, we employ both height-wise and width-wise partitions in our model for vehicle Re-ID. Moreover, filters in a convolutional layer generate channel information. Even though inputs are the same, they learn and update their parameters independently. Local features on channel-wise partitioned parts can be different from global features. By 3D partitions, our proposed PRN is able to extract maximally distinct local features, which helps to build robust vehicle appearance signatures.

3.3. Loss function

Softmax Cross-entropy loss. Softmax Cross-entropy loss is the most common used loss function in image classification tasks. The loss in a mini-batch can be described as:

$$L_{softmax} = -\sum_{i=1}^{N_i} \log\left(\frac{\exp(x[y])}{\sum_{j=1}^{N_{id}} \exp(x[j])}\right) \quad (1)$$

where N_i denotes the number of images in the mini-batch, N_{id} is the number of identities in the whole training set. y is the ground truth identity of input image and $x[j]$ represents the output of fully-connected layer for jth identity.

Triplet loss. In a mini-batch which contains P identities and K images for each identity, each image (anchor) has $K-1$ images of same identity (positives) and $(P-1)*K$ images of different identities (negatives). Triplet loss aims at pulling the positive pair (a, p) together while pushing the negative pair (a, n) away by a margin. Batch hard triplet loss is introduced in [10] as a variant of traditional triplet loss, which aims at laying more weight on most closest negative and farthest positive pairs. The batch hard triplet loss can be defined as:

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K}[\max_{p=1,\dots,K} \|\mathbf{a_i} - \mathbf{p_i}\|_2 - \min_{\substack{n=1,\dots,K \\ j=1,\dots,P \\ j\neq i}} \|\mathbf{a_i} - \mathbf{n_j}\|_2 + \alpha]_+ \quad (2)$$

where $\mathbf{a_i}$, $\mathbf{p_i}$ and $\mathbf{n_i}$ are the feature vectors of anchor, positive and negative samples respectively, and α is the margin to control the difference between positive and negative pair distances.

Total loss. The total loss is used to train our proposed method in an end-to-end manner, which combines softmax cross entropy losses with triplet losses.

$$L_{total} = \lambda \frac{1}{N_{CE}} \sum_{i=1}^{N_{CE}} L_{CE} + \frac{1}{N_{triplet}} \sum_{i=1}^{N_{triplet}} L_{triplet} \quad (3)$$

where $N_{softmax}$ and $N_{triplet}$ are the number of softmax cross entropy losses and triplet losses respectively. In our proposed PRN, all 18 outputs of fully connected layers are fed into softmax cross entropy loss and 2 global feature vectors are trained with triplet loss, so we have $N_{CE} = 18$ and $N_{triplet} = 2$. Parameter λ balances the contribution of two kinds of loss functions.

4. Experiment Details

4.1. Datasets and Evaluation metrics

Extensive experiments are conducted on four large-scale vehicle Re-ID datasets, *i.e.*, VeRi-776, VehicleID, VRIC and CityFlow-ReID to validate the performance of our proposed PRN.

The VeRi-776 dataset [21] is divided into two subsets for training and testing. The training set contains 37,781 images of 576 vehicles and the testing set contains 11,579 images of 200 vehicles. Following the evaluation protocol in [21], the image-to-track cross-camera search is performed, which means we use one image of a vehicle from one camera to search for tracks of the same vehicle in other cameras. We use mean average precision (mAP), as well as Top-1 and Top-5 accuracy of cumulative match curve (CMC).

The VehicleID dataset [19] is much larger than VeRi-776. It contains data captured during daytime by multiple real-world surveillance cameras distributed in a small city in China. There are 26,267 vehicles (221,763 images in total) in the entire dataset. The training set has 113,346 images of 13,164 vehicles and the rest is used for testing. Along with the dataset, 3 test protocols of different size, shown in Table 1, are proposed in the original paper. Since one camera only takes one image for same vehicle, we neglect cross-camera setting and conduct image-to-image search for testing. For each test protocol, one image of each identity is randomly selected to form a gallery set, and all the remaining images are used as the query set. The average results of 10-time-repeated random sampling are reported as the performance of our model on VehicleID dataset. Evaluation metrics for VehicleID dataset are Top-1 and Top-5 accuracy of CMC.

Protocol	Small	Medium	Large
Identity size	800	1,600	2,400
Image size	6,493	13,377	19,777

Table 1. Three test protocols of VehicleID dataset.

The VRIC dataset [15] is a newer dataset, which consists of 60,430 images of 5,656 vehicle IDs collected from 60 different cameras in traffic scenes. VRIC differs significantly from existing datasets in that unconstrained vehicle appearances were captured with variations in imaging resolution, motion blur, weather condition, and occlusion. The training set has 54,808 images of 2,811 vehicles, while the rest 5,622 images of 2,811 identities are used for testing. Evaluation metrics for VehicleID dataset are Top-1 and Top-5 accuracy of CMC.

The CityFlow-ReID dataset [26] is used as the evaluation protocol at the 2019 AI City Challenge. In total, it contains 56,277 bounding boxes, where 36,935 of them belonging to 333 vehicle identities in the training set. The test set consists of 18,290 bounding boxes belonging to the other 333 identities. The rest of the 1,052 images are the queries. On average, each vehicle has 84.50 image signatures from 4.55 camera views. Camera information is not available, so cross-camera search is not necessary. Train and test track information is present, but absent for query images. Therefore, both image-to-image and image-to-tracklet searches are possible. To adopt the image-to-tracklet matching, signatures of each image in a tracklet are aggregated together by a mean pooling to build an overall tracklet signature.

4.2. Implementation details

All the input images are resized to $384 * 384$ so that enough information is maintained on each partitioned part. We use a ResNet-50 pretrained on ImageNet [5] as our backbone network to accelerate the training process. In the backbone network, we set the stride of down-sampling in the last convolutional layer to 1 and duplicate all the layers after conv4_1 into 2 independent branches. Each 1*1 convolutional layer is followed by a Batch Normalization layer and a fully connected layer. All the 1*1 convolutional layers, batch normalization layers and fully connected layers do not share weights. To make our proposed model more robust, we apply a Random Horizontal Flip and a Random Erasing [33] of a probability of 0.5 for data augmentation. The batch size is set to 16 with randomly selected 4 identities and 4 images for each identity. We train our model with

Architecture	Top-1	mAP
B(w/o partition)	93.09	72.76
B(HP)	91.95	74.54
B(WP)	93.33	76.91
B(CP)	95.53	79.09
B(HP+WP)	94.76	80.56
B(HP)+B(WP)	95.59	82.92
B(HP+CP)	96.25	82.85
B(WP+CP)	96.25	83.04
B(HP+WP+CP)	96.54	84.26
B(HP+CP)+B(WP+CP)	**97.14**	**85.84**

Table 2. Performance comparison of different architectures of PRN on VeRi-776 dataset at 2 evaluation metrics: Top-1 and mAP where the bold font denotes the best performance. "B" refers to an independent branch. "HP", "WP" and "CP" refer respectively to height-wise, width-wise and channel-wise partition.

L_{total}	Top-1	mAP
w/o $L_{triplet}$	96.66	84.90
$\lambda = 1$	96.42	84.52
$\lambda = 2$	97.14	**85.84**
$\lambda = 3$	**97.26**	85.79

Table 3. Performance comparison of training PRN with different L_{total} on VeRi-776 dataset at 2 evaluation metrics: Top-1 and mAP where the bold font denotes the best performance.

an AMSGrad [23] based Adam optimizer for 500 epochs. The weight decay factor for L2 regularization is set to 5e-4. The initial learning rate is set to 2e-4. It decays to 2e-5 after 300 epochs and to 2e-6 after 400 epochs. The margin α in triplet loss is set to 1.2 in all experiments and the parameter λ in total loss is set to 2. During evaluation, we concatenate all the feature vectors after 1*1 convolutional layer together as appearance representations for each image in query and gallery sets. Both feature representations extracted from original and horizontally flipped image are summed up and normalized to be the final vehicle appearance signature of an input image. Our model is implemented on PyTorch framework and takes about 6 hours to train on a single NVIDIA 1080 Ti GPU for VeRi-776 dataset.

4.3. Component Analysis

Extensive experiments are conducted on VeRi-776 dataset to verify the effectiveness of crucial components in PRN. We compare performance of different structure and loss function variants to find the optimal architecture for our proposed PRN.

Partition and Structure: we mainly compare different combinations of height-wise, width-wise and channel-wise partitions. We report performance of architecture variants in Table 2. The results show that partitions can generally improve the performance of neural networks in vehicle ReID tasks. Among 3 types of partition, channel-wise partition brings the highest improvement, while width-wise partition brings the least improvement. Combination of different types of partition yields better performance than single partition. In addition, adopting height-wise and width-wise partition separately in 2 branches strengthens independence of learned spatial features, which further boosts the performance of PRN.

Loss function: Softmax loss is the most common loss function in multi-class image classification tasks. Triplet loss makes it possible to lay more weights on hard samples. To balance contributions of 2 loss functions on hard and normal samples, a weight parameter λ should be determined. We test several possibilities of L_{total}, such as no triplet loss, equal weights on softmax loss and triplet loss, and more weights on softmax loss. In Table 3, results show that triplet loss slightly increase the performance of our model. But more weights are supposed to be laid on softmax loss. In the end, we set $\lambda = 2$ for all experiments.

4.4. Comparison with State-of-the-art

We compare our proposed model PRN with state-of-the-art methods on the 4 datasets, *i.e.*, VeRi-776, VehicleID, VRIC and CityFlow-ReID with corresponding evaluation metrics.

VeRi-776: Table 4 presents the result comparison between previous state-of-the-art and our model on VeRi-776 dataset. Our proposed model PRN achieves 97.14% on Top-1 accuracy, 99.4% on Top-5 accuracy and 85.84% on mAP without re-ranking [32]. These results surpass previous state-of-the-art on all the 3 metrics, especially mAP. Re-ranking can further enhance the performance of PRN. A good mAP score demonstrates that PRN has a stronger capacity to retrieve all the corresponding images of same identity in the gallery set, regardless of different camera properties and viewpoint changes.

VehicleID: The comparison of results on Vehicle-ID dataset is reported in Table 5. Since all the images are captured by cameras placed either in front or in back of vehicles, only front and rear views of vehicle are present in Vehicle-ID dataset. Side views of vehicle are absent. Due to the symmetry along width dimension in both front and rear views of vehicle, partition along width dimension makes PRN consider too much duplicated information. Hence, besides results of our proposed PRN, we also report a simplified version that with only Height-Channel Branch. Results show that RNN-HA(ResNet+672) [29] still has the best performance on Top-1 accuracy. But our simplified PRN outperforms previous state-of-the-art on Top-5 accuracy.

VRIC: VRIC is a relatively new-released dataset, in consequence, only few results have been reported as previous

Method	Top-1	Top-5	mAP	Publication	Annotation
FACT [20]	59.65	75.27	19.92	ICME'16	ID + Attr
FACT + Plate-SNN + STR[21]	61.44	78.78	27.77	ECCV'16	ID + Plate + ST
Siamese-CNN+Path-LSTM [24]	83.48	90.04	58.27	ICCV'17	ID + ST
OIFE [28]	89.43	-	48.00	ICCV'17	ID + KP
VAMI [35]	77.03	90.82	50.13	CVPR'18	ID + Attr + VP
VAMI [35] + STR [21]	85.92	91.84	61.32	CVPR'18	ID + Attr + VP + ST
RNN-HA(ResNet) [29]	80.79	92.31	56.80	ACCV'18	ID+Attr
GS-TRE [2]	96.24	98.97	59.47	IEEE Trans. Multimed.18	ID
PRN(ours)	97.14	**99.40**	85.84		ID
PRN(ours) + RR	**97.38**	98.87	**90.48**		ID

Table 4. Comparison of results (%) on VeRi-776 dataset with 3 evaluation metrics: Top-1, Top-5 and mAP where the bold font denotes the best method. ID refers to identity labels. Attr refers to Attributes annotations, such as color and model. Plate refers to extra plate number datasets. KP refers to Key points. VP refers to Viewpoint labels. ST stands for spatio-temporal information. RR stands for Re-Ranking [32].

Setting	Test Size=800		Test Size=1600		Test Size=2400	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
FACT [21]	49.53	68.07	44.59	64.57	39.92	60.32
Mixed Diff+CCL [19]	48.93	75.65	45.05	68.85	41.05	63.38
XVGAN [34]	52.87	80.83	49.55	71.39	44.89	66.65
VAMI [35]	63.12	83.25	52.87	75.12	47.34	70.29
GS-TRE [2]	75.9	84.2	74.8	83.6	74.0	82.7
RNN-HA(ResNet+672) [29]	**83.8**	88.1	**81.9**	87.0	**81.1**	87.4
PRN(ours)	63.07	89.29	55.42	84.23	50.36	79.14
PRN(Single Height-Channel Branch)	78.92	**94.81**	74.94	**92.02**	71.58	**88.46**

Table 5. Comparison of results (%) on VehicleID dataset with 2 evaluation metrics: Top-1 and Top-5 where the bold font denotes the best method.

L_{total}	Top-1	Top-5
Siamese-Visual [24]*	30.55	57.30
OIFE(Single Branch) [28]*	24.62	50.98
MSVF [15]	46.61	65.58
PRN(ours)	**80.42**	**94.83**

Table 6. Comparison of results (%) on VRIC dataset with 2 evaluation metrics: Top-1 and Top-5 where the bold font denotes the best method. * refers to results represented in [15].

Method	Top-1	mAP
DenseNet121+Xent+Htri [26]	51.7	31.0
ResNext101+Xent+Htri [26]	48.8	32.0
MobileNetV1+BH [26]	48.4	32.0
PRN(ours)-I2I+RR	59.89	42.75
PRN(ours)-I2T	62.17	49.48
PRN(ours)-I2T+Fusion	**65.97**	**53.44**

Table 7. Comparison of results (%) on CityFlow-ReID dataset with 2 evaluation metrics: Top-1 and mAP. I2I and I2T refer respectively to image-to-image and image-to-tracklet search. RR stands for Re-Ranking [32]. Fusion refers to multiple backbone network feature fusion.

work. We compare the results of our proposed PRN with several results mentioned in the paper of VRIC dataset [15]. As shown in Table 6, our PRN outperforms previous state-of-the-art on both Top-1 and Top-5 accuracy by a large margin.

CityFlow-ReID: We compare the results of our proposed PRN with several baselines mentioned in the paper of CityFlow dataset [26]. Results are reported in Table 7. Image-to-tracklet search shows a superior performance compared to image-to-image search. Our PRN outperforms these baselines on both Top-1 and accuracy and mAP by a large margin. To get a better performance, we also changed the backbone network from ResNet-50 to ResNet-152, DenseNet-161 [13] and SeNet-152 [12]. Final appearance signature for a vehicle is the average of appearance signatures built on these 3 deeper backbone networks. Our fusion based PRN achieved 13th place among 84 teams in the 2019 AI CITY CHALLENGE [1] track 2 vehicle Re-ID task.

5. Conclusion

In this paper, we focus on the well-considered partition strategies for person Re-ID and adapt it for vehicle Re-ID task. By conducting partitions along each dimension in the feature map, more local features are extracted to complement global features. A 2-branch structure is further proposed to reduce the number of duplicated features extracted from spatial dimensions (height and width). We conduct extensive ablation studies on VeRi-776 dataset to verify the effectiveness of each component. Without using any extra annotation, except identities, we propose a novel end-to-end trainable model called PRN for vehicle Re-ID, which outperforms current state-of-the-art.

References

[1] AI CITY CHALLENGE. Challenge track 2: City-scale multi-camera vehicle re-identification. https://www.aicitychallenge.org/, 2019. 7

[2] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. yu Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 20:2385–2399, 2018. 3, 7

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *CoRR*, abs/1405.3531, 2014. 3

[4] 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. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1335–1344, 2016. 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, 2010. 3

[7] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. S. Huang. Horizontal pyramid matching for person re-identification. *CoRR*, abs/1804.05275, 2018. 3

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 3

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[10] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. 2, 4

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 3

[12] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 7

[13] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 7

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[15] A. Kanaci, X. Zhu, and S. Gong. Vehicle re-identification in context. In *Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, September 10-12, 2018, Proceedings*, 2018. 5, 7

[16] F. M. Khan and F. Brémond. Multi-shot person re-identification using part appearance mixture. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 605–614, 2017. 3

[17] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. *CoRR*, abs/1901.01015, 2019. 3

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 2

[19] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2167–2175, 2016. 3, 5, 7

[20] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle re-identification in urban surveillance videos. *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016. 1, 3, 7

[21] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV 2016*, 2016. 3, 5, 7

[22] X. Liu, H. Ma, H. Fu, and M. Zhou. Vehicle retrieval and trajectory inference in urban traffic surveillance scene. In *ICDSC*, 2014. 3

[23] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. 6

[24] 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1918–1927, 2017. 3, 7

[25] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 3

[26] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. C. Anastasiu, and J.-N. Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *CVPR 2019: IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 5, 7

[27] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia*, 2018. 3

[28] 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 379–387, 2017. 7

[29] X.-S. Wei, C.-L. Zhang, L. Liu, C. Shen, and J. Wu. Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification. In *Asian Conference on Computer Vision (ACCV'18)*, 2018. 6, 7

[30] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 2019. 3

[31] D. Zapletal and A. Herout. Vehicle re-identification for automatic video traffic surveillance. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1568–1574, 2016. 3

[32] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017. 6, 7

[33] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017. 5

[34] Y. Zhou and L. Shao. Cross-view gan based vehicle generation for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. 7

[35] Y. Zhou and L. Shao. Viewpoint-aware attentive multi-view inference for vehicle re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2018. 3, 7