Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification

Xinchen Liu[1], Wu Liu[1*], Jinkai Zheng[2], Chenggang Yan[2], Tao Mei[1]
[1]AI Research of JD.com, [2]Hangzhou Dianzi University
{liuxinchen1,liuwu1}@jd.com,{zhengjinkai3,cgyan}@hdu.edu.cn,tmei@jd.com

ABSTRACT

Vehicle re-identification (Re-Id) is a challenging task due to the inter-class similarity, the intra-class difference, and the cross-view misalignment of vehicle parts. Although recent methods achieve great improvement by learning detailed features from keypoints or bounding boxes of parts, vehicle Re-Id is still far from being solved. Different from existing methods, we propose a Parsing-guided Cross-part Reasoning Network, named as PCRNet, for vehicle Re-Id. The PCRNet explores vehicle parsing to learn discriminative part-level features, model the correlation among vehicle parts, and achieve precise part alignment for vehicle Re-Id. To accurately segment vehicle parts, we first build a large-scale Multi-grained Vehicle Parsing (MVP) dataset from surveillance images. With the parsed parts, we extract regional features for each part and build a part-neighboring graph to explicitly model the correlation among parts. Then, the graph convolutional networks (GCNs) are adopted to propagate local information among parts, which can discover the most effective local features of varied viewpoints. Moreover, we propose a self-supervised part prediction loss to make the GCNs generate features of invisible parts from visible parts under different viewpoints. By this means, the same vehicle from different viewpoints can be matched with the well-aligned and robust feature representations. Through extensive experiments, our PCRNet significantly outperforms the state-of-the-art methods on three large-scale vehicle Re-Id datasets.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Image segmentation**; **Object identification**.

KEYWORDS

Vehicle Re-identification, Vehicle Parsing, Image Segmentation, Graph Convolutional Network, Self-supervised Learning

ACM Reference Format:
Xinchen Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, Tao Mei. 2020. Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20), October 12-16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413578

*Wu Liu is the corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA.

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413578

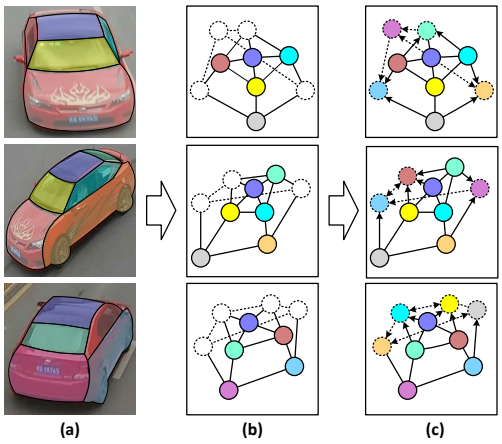

Figure 1: (a) A vehicle from varied views with parsed masks. (b) Part-neighboring graphs for part alignment. (c) Invisible feature prediction from visible parts. (Best viewed in color.)

1 INTRODUCTION

Vehicle re-identification (Re-Id) is, given a probe vehicle image, to search for the same vehicle captured by large-scale camera networks. Due to its wide applications such as intelligent transportation and public security, it attracts tremendous attention from the multimedia and computer vision communities [20]. Existing methods achieves great improvement with large-scale data and deep convolutional neural networks (CNNs) [24, 31].

However, compared with person Re-Id [15, 42], vehicle Re-Id has unique characteristics and faces specific challenges that make vehicle Re-Id far from being solved. One challenge is the trivial inter-class differences between different vehicles of similar viewpoints, especially for the vehicles of the same model and color. Therefore, it difficult to distinguish vehicles only based on the global visual features, while the precise local details of vehicle parts must be exploited for vehicle Re-Id, as shown in Figure 1 (a). Moreover, since the vehicle has a rigid body, vehicle Re-Id is a type of multi-view matching problem, which makes the visible parts of one vehicle captured from various viewpoints extremely different. This brings ambiguity for the alignment of vehicle parts during the matching of part-level features. These challenges motivate us to discover the correlation among vehicle parts to predict invisible parts from visible parts and learn more robust representation for vehicle Re-Id, as shown in Figure 1 (b) and (c).

Existing vehicle Re-Id methods develop in two stages. Early studies mainly focus on the global appearance and exploit metric learning methods to learn an embedding space, in which samples of the same vehicle are close while those of different vehicles are distant [7, 18]. However, due to the ambiguous appearances of vehicles under varied viewpoints, the metric learning model cannot obtain optimal results with only global features. Recent works use more detailed annotations, such as keypoints and viewpoints, to discover local representations for vehicle Re-Id [32, 44]. Most recently, He *et al.* propose to locate the bounding boxes of vehicle parts and integrate both regional and global features, which achieves excellent performance [8]. However, viewpoints, keypoints, and bounding boxes only provide coarse local information. Moreover, these methods consider keypoints or regions separately and neglect the relations among different parts for vehicle representation. Therefore, we explore vehicle parsing to learn discriminative local features and then discover the relations among parts to achieve invisible part prediction.

To achieve accurate vehicle parsing and facilitate Re-Id as well as other related vehicle analysis tasks in real-world applications, we build a Multi-grained Vehicle Parsing (MVP) dataset. The MVP dataset has the following featured properties: 1) It contains 24,000 vehicle images of varied resolutions captured in real surveillance scenes. 2) We annotate two granularities of pixel-level vehicle parts for different requirements, i.e., the ten coarse-grained classes and the 59 fine-grained classes. 3) The images are collected from several public vehicle Re-Id datasets which not only cover diverse vehicle types, models, and colors, but also reflect variations of surveillance scenes, such as viewpoints, illumination, and backgrounds. In addition to vehicle Re-Id, the MVP dataset can be used in many potential applications such as vehicle categorization [36, 37], tracking [4], retrieval [33], and autonomous driving [5].

With the parsed vehicle parts, we propose a Parsing-guided Cross-part Reasoning Network (PCRNet) to learn discriminative feature representations and model the correlation among parts for vehicle Re-Id. The PCRNet explores the cross-part relations through two well-designed branches. Based on the structure of vehicles, we first build a part-neighboring graph for part correlation mining. Then, one branch of PCRNet adopts the graph convolutional networks (GCNs) to perform information propagation among vehicle parts. By this means, the correlation between neighboring parts can be modeled and the significant parts from a specific viewpoint will be highlighted. Moreover, we propose a self-supervised part prediction loss that can enable the GCNs to predict features of invisible parts based on those of visible parts. Furthermore, the other branch learns global features with a novel parsing-based part erasing augmentation, which makes the model robust to invisible parts. By integrating the two complementary branches, the PCRNet learns a comprehensive representation for vehicle Re-Id. In summary, the contributions of this paper include:

- We build the first large-scale Multi-grained Vehicle Parsing dataset for real-world surveillance scenes. We also provide comprehensive evaluation of the state-of-the-art semantic segmentation methods on the MVP dataset.
- We propose a Parsing-guided Cross-part Reasoning Network for vehicle Re-Id. The PCRNet exploits vehicle parsing to extract part-level features and explicitly align vehicle parts for the discriminative feature representation.
- We adopt the GCNs with an elaborate self-supervised part prediction loss to discover the cross-part correlation and generate features of invisible parts from visible parts.

Through extensive experiments, the proposed PCRNet outperforms the state-of-the-art methods on three large-scale benchmarks, i.e., VeRi [20], VehicleID [18], and VeRi-Wild [24].

2 RELATED WORK

Vehicle Re-Identification. Vehicle Re-Id methods can be categorized into two classes: vision-based methods that only utilize the content of vehicle images [18, 30] and multi-modal methods that adopt information of other modalities such as license plates or spatiotemporal context of surveillance networks [19, 27, 32]. This paper mainly focuses on vision-based vehicle Re-Id. Early studies usually take the whole images as the input and adopt the metric learning-based model to learn a discriminative latent space for vehicle Re-Id [7, 18]. In recent works, researchers explore local information such as keypoints, viewpoints, and parts to capture local details for vehicle representations [8, 30, 32, 44]. For example, Wang *et al.* labeled 20 keypoints of vehicles to extract local features for vehicle Re-Id [32]. Zhou *et al.* proposed a viewpoint-aware attentive multi-view inference framework to capture shared regional features in different viewpoints [44]. Tang *et al.* proposed to explicitly reason about vehicle pose and shape via keypoints, heatmaps, and segments for vehicle Re-Id [30]. He *et al.* utilized a detection model to local several key parts and combined global and regional features, which obtained the state-of-the-art results for vehicle Re-Id [8]. However, since viewpoints, keypoints, and bounding boxes only provide coarse and limited local information, more discriminative details may be neglected. In addition, these methods usually integrate local features by direct concatenation but neglect their relations for the representation of vehicles. Therefore, this paper explores vehicle parsing and delves into cross-part correlation for vehicle Re-Id.

Fine-grained Image Parsing. Fine-grained image parsing is a specific task of semantic segmentation for pixel-level classification of object parts in images. Existing methods and datasets mainly focus on clothing parsing [34, 35], human parsing [6, 14, 16, 40], and face parsing [22, 28]. For example, Yamaguchi *et al.* built a large-scale Fashionista dataset and pioneered early research on clothing parsing [34, 35]. Gong *et al.* released the first large-scale human parsing dataset and proposed to jointly model pose estimation and human parsing by a multi-task learning framework [16]. Liu *et al.* constructed a large-scale landmark guided face parsing dataset and proposed a boundary-attention semantic segmentation method for face parsing [22]. Although there are datasets containing labels of car part masks such as the 3D Object Class Dataset [26] and the PASCAL-Part Dataset [3], few large-scale datasets are dedicated to fine-grained vehicle parsing in the wild. Therefore, we construct a large-scale Multi-grained Vehicle Parsing dataset, which can be used not only for vehicle Re-Id but also for other vehicle analysis tasks. Moreover, human parsing has been adopted to improve the performance of person Re-Id [10], which also inspires us to explore vehicle parsing for vehicle Re-Id.

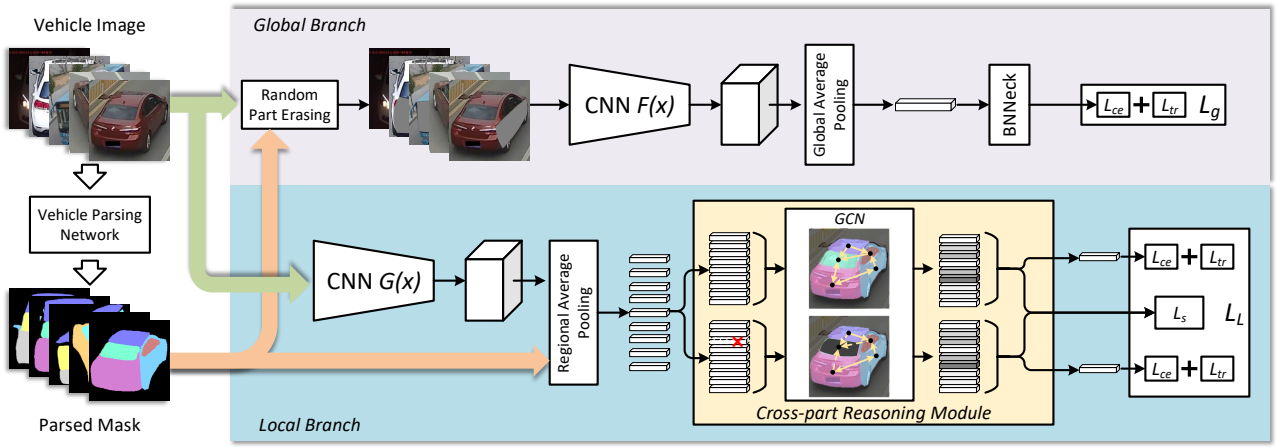

Figure 2: The overall architecture of the Parsing-guided Cross-part Reasoning framework. (Best viewed in color.)

3 PCRNET FOR VEHICLE RE-ID

3.1 Overview

As discussed in previous sections, different vehicles usually have very similar global appearances and must be distinguished by detailed local features. Besides, the same vehicle may be captured from different viewpoints, which makes alignment of part-level features difficult. Therefore, this section introduces a Parsing-guided Cross-part Reasoning Network, named as PCRNet, for vehicle Re-Id, as shown in Figure 2. In our framework, the PCRNet directly uses an image segmentation model trained on vehicle parsing data as a preprocessing tool to obtain the parsed masks of vehicle images. The PCRNet has two branches for learning local details and global features, respectively. In the local branch, a parsing-guided cross-part reasoning module is designed to discover the correlation among vehicle parts. This module builds a part-neighboring graph based on the structure of the vehicle body. Then the graph convolutional network (GCN) takes the graph and the regional features of vehicle parts as the input to perform feature propagation among parts. By this means, the vehicle parts can be explicitly aligned and the discriminative local features can be enhanced. Moreover, we design a self-supervised learning scheme to guide the GCN learn to generate features of invisible parts from visible parts. The global branch takes the vehicle image as a whole to learn the global appearance of vehicle. With masks of vehicle parts, we design a random part erasing augmentation method, which makes the learned global features more robust to varied viewpoints and occlusion. At last, the two branches are jointly optimized in an end-to-end manner.

3.2 Cross-part Reasoning with GCN

To discover discriminative local features and model the cross-part correlation for vehicle Re-Id, we design the local branch with two main components: 1) parsing-guided regional feature extraction and 2) GCN-based cross-part reasoning.

Regional Feature Extraction. The local branch adopts a deep convolutional neural network (CNN), i.e., ResNet [9], with a Batch Normalization Neck (BNNeck) [25] as the backbone, since it demonstrates the powerful capability of representation learning for person Re-Id. By feeding the input $\mathbf{x}$ into the CNN $G(\cdot)$, a semantic feature map $G_m(\mathbf{x}) \in \mathbb{R}^{l\times w\times h}$ is obtained from the last convolutional layer. Then the local branch performs regional average pooling (RAP) to obtain a feature vector for each part. Before RAP, we first resize the parsed mask, $M \in \mathbb{N}^{W\times H}$, to $\widetilde{M} \in \mathbb{N}^{w\times h}$ of the same size with the feature map. Then the regional feature map $G_c(\mathbf{x})$ for part c is calculated as:

$$\begin{aligned} G_c(\mathbf{x}) &= G_m(\mathbf{x}) \odot \widetilde{M^*}, \\ \widetilde{M^*}_{i,j} &= \mathcal{I}_c(\widetilde{M}_{i,j}), \end{aligned} \tag{1}$$

where $\odot$ is the element-wise multiplication, $\mathcal{I}_c(\cdot)$ is an indicator function that returns 1 if its input equals to c and 0 otherwise. After that, we average the non-zero elements on each channel of $G_c(\mathbf{x})$ to obtain the regional feature vector $\mathbf{r}_c \in \mathbb{R}^l$ for part c, where l is the channel number of the last convolutional layer.

GCN-based Cross-part Reasoning. In existing methods, local features of keypoints or bounding boxes are usually fused by direct concatenation [8, 32], which neglects the relations among vehicle parts. Meanwhile, the vehicle parts involve natural neighboring relations such as (roof, front-windshield), (left-window, left-body), and so on. Therefore, we build a part-neighboring Graph (PNG) to explicitly model these relations, as shown in Figure 1. The PNG can be formulated by an adjacent matrix $A \in \mathbb{R}^{C\times C}$, where C is the number of nodes, i.e., the pre-defined parts ($C = 9$ in our implementation). For the adjacent matrix A, we set $A(i, j) = 1$ if part i and part j are neighboring. To effectively mine discriminative local features with PNG, we adopt the GCN to perform relational reasoning by information propagation from each node to its neighbors in the graph [13, 38]. At the tail of backbone network, we add a two-layer GCN, in which each layer L is formulated as:

$$X^{(L)} = \sigma(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}X^{(L-1)}W^{(L)}), \tag{2}$$

where $A \in \mathbb{R}^{C\times C}$ is the adjacent matrix, $D \in \mathbb{R}^{C\times C}$ is the degree matrix of A, $X^{(L-1)} \in \mathbb{R}^{C\times l}$ is the output feature matrix of the $L-1$-th layer, $W^{(L)}\mathbb{R}^{l\times l}$ is the learnable parameters of layer L, and $\sigma(\cdot)$ is an activation function. The initial feature matrix $X^{(0)}$ is obtained by the regional features, i.e., $X^{(0)} = [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_C]^T$. The outputs of the GCN, $X^{(L)}$, are the updated features by L-round message

propagations among the nodes in PNG. At last, $X^{(L)} \in \mathbb{R}^{C\times l}$ is averaged over all nodes to obtain the final local feature, $\mathbf{r} \in \mathbb{R}^l$.

Self-supervised Part Prediction Loss. To make the GCN able to predict the features of invisible parts based on visible parts, we propose a self-supervised learning strategy for the local branch, as shown in Figure 2. Before feeding the initial feature matrix $X^{(0)}$ into the GCN, we randomly set a row of feature vector to zero for imitation of an invisible part and obtain a new feature matrix $\hat{X}^{(0)}$. In each iteration, $X^{(0)}$ and $\hat{X}^{(0)}$ are fed into the same GCN to obtain the local features $\mathbf{r}$ and $\hat{\mathbf{r}}$. At last, the loss function of the local branch can be formulated as:

$$\mathcal{L}_l = \lambda_1(L_t(\mathbf{r}) + L_c(\mathbf{r})) + \lambda_2(L_t(\hat{\mathbf{r}}) + L_c(\hat{\mathbf{r}})) + \alpha||\mathbf{r} - \hat{\mathbf{r}}||_2, \quad (3)$$

where $L_t(\cdot)$ and $L_c(\cdot)$ are the triplet loss and the cross entropy loss, respectively, the last item is the Euclidean distance between $\mathbf{r}$ and $\hat{\mathbf{r}}$, and $\lambda_1, \lambda_2, \alpha$ are hyper-parameters to balance the losses.[1] Through GCN-based reasoning with self-supervised learning, the cross-part correlation can be discovered to generate a more discriminative local representation for vehicle Re-Id.

3.3 Global Appearance Learning

Notwithstanding the effectiveness of local features, global appearances such as shape, type, color, and model are also of great importance for vehicle Re-Id [7, 18]. For example, one can easily distinguish a yellow car from a red car or an SUV from a truck by just observing their overall appearances even they are captured from different viewpoints or with occlusion.

In the global branch of the PCRNet, we also adopt the ResNet [9] with a BNNeck [25] as the backbone. Moreover, to overcome the viewpoint variation and occlusion, we propose a parsing-guided random part erasing method as a data augmentation scheme for the global branch. For person Re-Id, random erasing (RE) [43] has been widely adopted to enhance the generalization of deep CNN models. However, as a general data augmentation, RE only provides a relaxed regularization on training the model for vehicle Re-Id. Therefore, we apply a stronger constraint on vehicle parts to perform a random part erasing augmentation. Given an input vehicle image, $\mathbf{x} \in \mathbb{R}^{C\times W\times H}$, and its parsed mask, $M \in \mathbb{N}^{W\times H}$, we randomly erase a vehicle part given a probability $p \in [0, 1]$ as:

$$\begin{aligned} \mathbf{x}^* &= zM^* + \mathbf{x} \odot (\mathbf{1} - M^*), \\ M^*_{i,j} &= \mathcal{I}_c(M_{i,j}), \end{aligned} \quad (4)$$

where z is the value to replace the erased pixels ($z = 0$ in our implementation), $\odot$ is the element-wise multiplication, $\mathcal{I}_c(\cdot)$ is an indicator function that returns 1 if its input equals to c and 0 otherwise, c is a part class that is randomly selected to be erased from all part classes in $M \in \mathbb{N}^{W\times H}$.

At last, given the global network $F(\cdot)$, we use the output of the BNNeck layer as the global feature, $\mathbf{g} = F(\mathbf{x})$, to compute the triplet loss, $L_t(\mathbf{g})$, and cross entropy loss, $L_c(\mathbf{g})$, respectively. The loss function of the global branch, i.e., $\mathcal{L}_g$ is defined as:

$$\mathcal{L}_g = L_t(\mathbf{g}) + L_c(\mathbf{g}). \quad (5)$$

[1]Here we omit the triplet input for the triplet loss for simplicity.

3.4 Training and Inference

As discussed above, the PCRNet has two branches for learning the global and local representations for vehicle Re-Id. Before training and testing of the PCRNet, we first train a vehicle parsing network on our vehicle parsing dataset. Therefore, during training the PCRNet, we construct a batch of samples with both vehicle images and their parsed masks obtained from the vehicle parsing network. To calculate the triplet loss in Equation 3 and Equation 5, we randomly select N IDs and K samples per ID to build the triplets. At last, the PCRNet is trained in the end-to-end manner with the objective function as follows:

$$\mathcal{L} = \mathcal{L}_g + \beta\mathcal{L}_l, \quad (6)$$

where β is a balance parameter.

During testing, the query images and gallery images are first fed into the vehicle parsing network to obtain parsed masks. Then, the PCRNet takes the image and the mask as the input to obtain the global feature $\mathbf{g}$ and the local feature $\mathbf{r}$. Next, the similarity between each pair of query i and gallery j can be estimated by

$$s = \mu D(g_i, g_j) + (1 - \mu)D(r_i, r_j) \quad (7)$$

where μ is the fusion weight and $D(\cdot, \cdot)$ is a distance metric.

4 THE MVP DATASET

In this section, we present a novel large-scale dataset, Multi-grained Vehicle Parsing (MVP), for semantic analysis of vehicles in the wild, which has several featured properties. First of all, the MVP contains 24,000 vehicle images captured in read-world surveillance scenes, which makes it more scalable than existing datasets, as listed in Table 1. Moreover, for different requirements, we annotate the vehicle images with pixel-level part masks in two granularities, i.e., the coarse annotations of ten classes and the fine annotations of 59 classes. The former can be applied to object-level applications such as vehicle Re-Id, fine-grained classification, and pose estimation, while the latter can be explored for high-quality image generation and content manipulation. Furthermore, the images reflect complexity of real surveillance scenes, such as different viewpoints, illumination conditions, backgrounds, and etc. In addition, the vehicles have diverse countries, types, brands, models, and colors, which makes the dataset more diverse and challenging.

4.1 Vehicle Image Collection

To guarantee the diversity of vehicles and complexity of environments, we collect the images from three large-scale vehicle Re-Id datasets, i.e., VeRi [20], CityFlow-ReId [31], and VERI-Wild [24]. The VeRi dataset has 49,325 images of 775 vehicles captured by 20 cameras. The CityFlow-ReID has contains 56,277 images of 666 vehicles captured by 40 cameras. The VERI-Wild dataset contains 416,314 images of 40,671 vehicles captured by 174 cameras. Each image of these datasets contains one vehicle cropped from a video frame. From these datasets, we randomly sample 40,000 images larger than 256×256 as a pool for pixel-level part annotation.

4.2 Vehicle Parsing Annotation

The annotation is performed manually by annotators in two steps. In the first step, the annotators label 9 coarse parts of vehicles for about 30,000 images sampled from the pool. The 9 parts include *roof*,

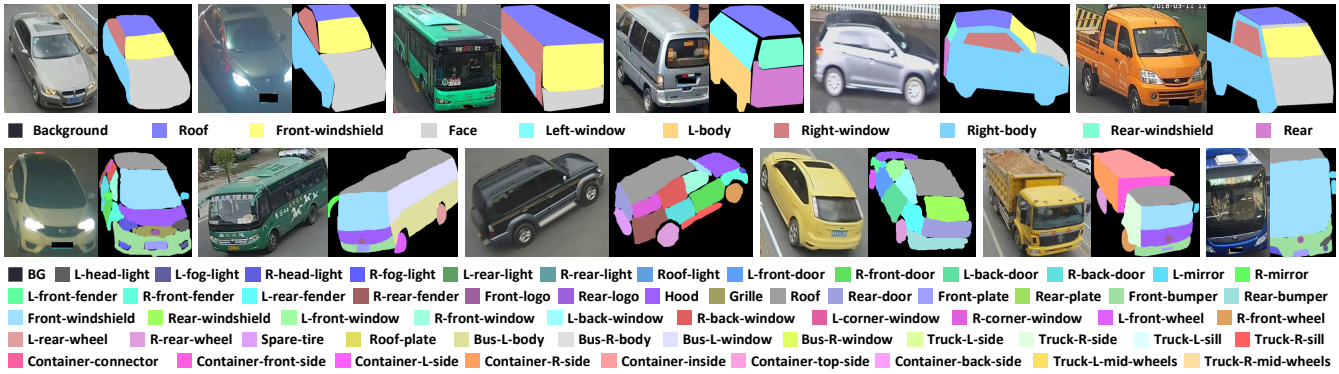

Figure 3: Some images and visualization of coarse and fine-grained parsing annotations. (Best viewed in color.)

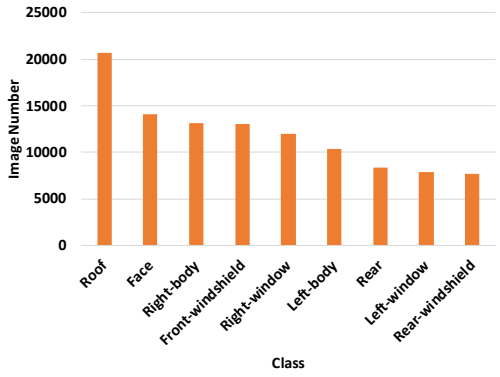

Figure 4: The image numbers over nine coarse parts.

Table 1: Comparison of public datasets for vehicle parsing.

Dataset	Class #	Image #	Surveillance?
3D Class Dataset-Car [26]	6	960	✗
Pascal-Part-Car [3]	14	1,805	✗
Pascal-Part-Bus [3]	14	501	✗
MVP-coarse	10	21,000	✓
MVP-fine	59	3,000	✓

front-windshield, *face*, *left-window*, *left-body*, *right-window*, *right-body*, *rear-windshield*, and *rear*. In the second step, for fine-grained annotation, we sample about 4,000 images from the rest images of the pool. These images are labeled with 58 fine-grained parts, including *left-head-light*, *left-fog-light*, *right-head-light*, *right-fog-light*, *left-rear-light*, *right-rear-light*, *roof-light*, *left-front-door*, *right-front-door*, *left-back-door*, *right-back-door*, *left-mirror*, *right-mirror*, *left-front-fender*, *right-front-fender*, *left-rear-fender*, *right-rear-fender*, *front-logo*, *rear-logo*, *hood*, *grille*, *roof*, *rear-door*, *front-plate*, *rear-plate*, *front-bumper*, *rear-bumper*, *front-windshield*, *rear-windshield*, *left-front-window*, *right-front-window*, *left-back-window*, *right-back-window*, *left-corner-window*, *right-corner-window*, *left-front-wheel*, *right-front-wheel*, *left-rear-wheel*, *right-rear-wheel*, *spare-tire*, *roof-plate*, *bus-left-body*, *bus-right-body*, *bus-left-window*, *bus-right-window*, *truck-left-side*, *truck-right-side*, *truck-left-sill*, *truck-right-sill*, *container-connector*, *container-front-side*, *container-left-side*, *container-right-side*, *container-inside*, *container-top-side*, *container-back-side*, *truck-left-mid-wheels*, and *truck-right-mid-wheels*. The non-vehicle region is labeled as the background.

During annotation, the annotators filter out the images with low quality, strong occlusions, and few parts. After annotation, all labels are inspected by two rounds to guarantee high quality. Finally, we obtain the MVP-coarse dataset containing 21,000 images with 9-part masks and the MVP-fine dataset of 3,000 images with 58-part masks. Some examples are shown in Figure 3.

4.3 Dataset Splitting and Statistics

To facilitate related research, we split the MVP-coarse dataset into the training/validation/testing subsets with 13K/4K/4K images, respectively. The MVP-fine dataset is split into the training/validation/testing subsets with 1,800/600/600 images, respectively. The resolution of images in MVP-coarse ranges from 256×256 to 2, 026×1, 174 with 458×388 on average. The resolution of images in MVP-fine ranges from 256×256 to 1, 517×1, 065 with 438×378 on average. The statistics of image numbers over coarse-grained parts and the fine-grained parts are shown in Figure 4 and the Appendix, respectively. We can find that the coarse-grained annotations have a relatively balanced distribution over nine parts. However, the fine-grained annotations show a long-tail distribution, in which 31 parts have over 1,000 samples while the rest parts have much fewer samples. This makes fine-grained segmentation more challenging than the coarse segmentation. In addition, the average numbers of parts per image for the coarse and the fine annotations are 5.1 and 17.0, respectively.

4.4 Empirical Study of Vehicle Parsing

In this section, we evaluate three state-of-the-art semantic segmentation methods proposed in recent years on our MVP dataset. The details of compared methods are as follows:

1) Pyramid Scene Parsing Network (PSPNet) [39]: The PSPNet is one of the state-of-the-art CNN-based models for semantic segmentation. It adopts a full-convolutional network with pyramid feature pooling for the multi-scale representation. We implement the PSPNet with ResNet-101 [9] as the backbone.

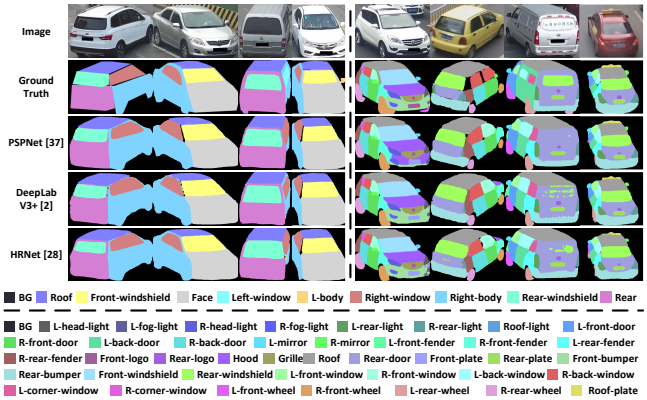

Figure 5: The visualized results on the validation set of MVP. (Best viewed in color with zoom-in.)

2) DeepLabV3+ [2]: This method is also one of the state-of-the-art semantic segmentation methods. It adopts the atrous convolution with different dilation rates to capture multi-scale features. We use the ResNet-101 [9] as its backbone for vehicle parsing.

3) High-Resolution Net (HRNet) [29]: The HRNet connects high-to-low resolution convolutions in parallel and repeatedly conducts multi-scale fusions across parallel convolutions. It achieves the state-of-the-art performance on semantic segmentation and human parsing. We adopt the HRNet-W32 as the backbone in our implementation.

For a fair comparison, we train these methods with the pixel-level cross-entropy loss and the same training strategy. Following human parsing [6], we use the pixel accuracy (Pixel Acc), mean accuracy (Mean Acc), and mean intersection over union (mIoU) as the evaluation metrics.

4.5 Evaluation on MVP

Quantitive results. The evaluation of the three methods for coarse and fine vehicle parsing are listed in Table 2 and Table 3. From the results, we can first observe that the three methods obtain competitive performance for coarse and fine vehicle parsing. Moreover, all methods achieve excellent results for coarse vehicle parsing. This means that the parsed results can be directly used in other tasks such as vehicle Re-Id. Therefore, we exploit the coarse parsed vehicle parts in our Parsing-guided Cross-part Reasoning framework. Furthermore, the performance of three methods on the fine-grained vehicle parsing task is much worse than the coarse parsing task. This reflects that fine-grained vehicle parsing is a more challenging task due to varied part appearance, unbalanced class distribution, and very small targets.

Qualitative evaluation. We visualize the parsed results of the three methods as shown in Figure 5. The results also show that, for coarse vehicle parsing, the compared methods perform well and can accurately segment vehicle parts under complex background and different viewpoints. For fine-grained vehicle parsing, large parts such as roof, doors, windows, and wheels can be well parsed. However, these models cannot work well for details, edges, and small parts like lights, logos, and mirrors.

Table 2: Results on the val (test) set of coarse annotations.

Method	Pixel Acc	Mean Acc	mIoU
PSPNet [39]	96.36 (90.26)	96.00 (89.08)	92.18 (79.78)
DeepLabV3+ [2]	96.93 (90.55)	96.69 (89.45)	93.49 (80.41)
HRNet [29]	95.49 (90.40)	95.01 (89.36)	90.37 (80.04)

Table 3: Results on the val (test) set of fine annotations.

Method	Pixel Acc	Mean Acc	mIoU
PSPNet [39]	86.56 (86.21)	71.10 (69.61)	58.36 (57.47)
DeepLabV3+ [2]	87.65 (87.42)	75.28 (73.50)	62.24 (61.60)
HRNet [29]	86.66 (86.47)	72.96 (72.62)	60.48 (60.21)

5 EXPERIMENTS

5.1 Datasets and Experimental Setting

In our experiment, we first compare our PCRNet with the state-of-the-art vehicle Re-Id methods on three widely-used datasets. Then we provide the ablation study on the VeRi dataset to demonstrate the effectiveness of each component in our framework. The details of the datasets are as follows.

The **VeRi** [20] dataset has 49,325 images of 775 vehicles are captured by 20 cameras with various viewpoints, complex backgrounds, and different distances. The dataset is split into a training set with 37,746 images of 575 IDs and a testing set with 11,579 images of 200 IDs. During testing, 1,678 images of the testing set are used as the queries, while the rest images are used as the gallery.

The **VehicleID** [18] dataset contains 221,763 images of 26,267 vehicles captured by 40 cameras. The images in VehicleID are captured by high-definition cameras and only have front and back views. Its training set contains 110,178 images of 13,134 vehicles, while the testing set contains 111,585 images of 13,133 vehicles. The authors of [18] extract three subsets from the testing set for the vehicle Re-Id task. The three subsets contain 800, 1,600, and 2,400 vehicles, respectively.

The **VERI-Wild** [24] dataset has 416,314 images of 40,671 vehicles captured by 174 cameras. The images not only involve complex backgrounds and various viewpoints, but also reflect various weather and illumination conditions. This dataset is randomly divided into a training set with 277,797 of 30,671 IDs and a testing set with 138,517 images of 10,000 IDs. The testing set also has three subsets containing 3,000, 5,000, and 10,000 IDs, respectively.

Following [20, 24], we use mean Average Precision (mAP), Rank-1 accuracy (R-1), and Rank-5 accuracy (R-5) as the evaluation metrics for vehicle Re-Id.

5.2 Implementation Details

This section presents the details of data preparation, the training strategy, and the testing process in our experiments.

Data Preparation. We adopt the HRNet [29] as the vehicle parsing network. The HRNet is trained on the MVP-coarse dataset with the input resolution of 384×384. Some examples and visualized parsing results on the three vehicle Re-Id datasets are shown in Figure 6. Before training our PCRNet, we generate the masks of all images of the three vehicle Re-Id datasets for efficiency.

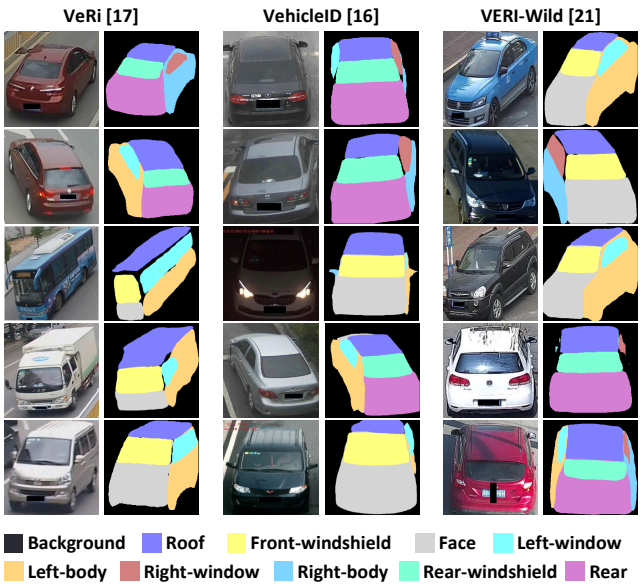

Figure 6: The visualized parsing results on three vehicle Re-Id datasets. (Best viewed in color with zoom-in.)

Networks Training. In our implementation, we adopt the ResNet-50 [9] model pre-trained on ImageNet as the backbone of the PCRNet. For all datasets, the input images and masks are resized to 256×256 with random horizontal flipping and random cropping as data augmentation. Our PCRNet is trained by the loss function in Equation 6. The hyper-parameters λ_1, λ_2, and α in Equation 2 are all set to 1.0. The β in Equation 6 is set to 0.5 for VeRi and 1.0 for VehicleID and VeRi-Wild. Inspired by [25], we adopt label smoothing for the cross entropy loss to alleviate overfitting. The model is optimized with the Adam optimizer [12] for 100 epochs. The learning rate is set to 3.5×10^{-4} and multiplied by 0.1 at epoch 40, 70, and 90. The warmup strategy is applied to the learning rate with the initial value 3.5×10^{-6} in the first 10 epochs.

Testing Setting. During testing, we use the official evaluation protocols of the three datasets. The μ in Equation 7 is set to 0.7 for VeRi and 0.6 for VehicleID and VeRi-Wild. We adopt the cosine distance as the $D(\cdot)$ in Equation 7.

5.3 Comparison of State-of-the-art Methods

5.3.1 Compared Methods. We compare our PCRNet with existing vehicle Re-Id methods which are categorized into four classes:

1) Hand-crafted feature-based methods. This class of methods includes LOMO [17] and BOW-CN [41] which represent the early studies before the rising of deep learning-based methods.

2) Global representation learned by deep CNNs. Representative approaches include GoogLeNet [37], Siamese-CNN [27], NuFACT [20], RAM [21], MLSL [1], and FDA-Net [24]. They adopt the deep CNNs to learn visual features from overall appearance.

3) Multi-modal methods. This type of methods usually exploits multi-modal information such as visual content, license plate, spatiotemporal context, and etc. OIFE+ST [32], Siamese+ST [27], and PROVID [20] are representative methods.

Table 4: Comparison of the state-of-the-art vehicle Re-Id methods on the VeRi dataset.

Methods	Year	mAP (%)	R-1 (%)	R-5 (%)
LOMO [17]	2015	9.6	25.3	46.5
BOW-CN [41]	2015	12.2	33.91	53.69
GoogLeNet [37]	2015	17.9	52.3	72.2
Siamese-CNN [27]	2017	54.2	79.3	88.9
NuFACT [20]	2018	48.5	76.9	91.4
RAM [21]	2018	61.5	88.6	94.0
FDA-Net [24]	2019	55.5	84.3	92.4
MLSL [1]	2019	61.1	90.0	96.0
OIFE+ST [32]	2017	51.4	92.4	-
Siamese-CNN+ST [27]	2017	58.3	83.5	90.0
PROVID [20]	2018	53.4	81.6	95.1
OIFE [32]	2017	48.0	65.9	-
VAMI [44]	2018	50.1	77.0	90.8
EALN [23]	2019	57.4	84.4	94.1
AAVER [11]	2019	61.2	89.0	94.7
PRN [8]	2019	70.2	92.2	97.9
PAMTRI [30]	2019	71.8	92.9	97.0
PRN*[8]	2019	74.3	94.3	**98.9**
PCRNet (ours)	2020	**78.6**	**95.4**	98.4

* Results with input size 512×512, others with 224×224 or 256×256.

4) Integration of global and local representations. Since our PCRNet explores vehicle parsing to learn effective representations for vehicle Re-Id, we mainly compare it with the state-of-the-art methods that also use auxiliary information such as keypoints, viewpoints, and bounding boxes of parts. The compared methods include OIFE [32], VAMI [44], C2F [7], EALN [23], AAVER [11], PRN [8], and PAMTRI [30].

5.3.2 Experimental Results. **Evaluation on VeRi.** The experimental results on the VeRi dataset are listed in Table 3. From the results, we can observe that most mainstream methods in the recent two years integrate local knowledge with global features for vehicle Re-Id, which improves the performance on the VeRi dataset by a large margin. For example, the mAP increases from about 50% $\sim$ 60% to more than 70%, while the Rank-1 accuracy improves from around 80% to over 90%. This demonstrates that local details are important cues for comprehensive and discriminative representations of vehicles. Because it is insufficient to distinguish two very similar vehicles only relying on the overall appearances such as shape, color, type, model, and etc. Moreover, for local features, regional annotations such as bounding boxes in PRN [8] or parsed mask in our PCRNet are more effective than keypoints or viewpoints in other methods. The reason is that regional knowledge, i.e., parts of vehicles, can make the deep CNNs focus on discriminative local regions such as lights, logos, grills, and even stuff in the vehicles seen through the windshield. Furthermore, our PCRNet outperforms the PRN method [8] and achieve the state-of-the-art performance on the VeRi dataset. The one reason is that parsed masks can provide more precise localization than bounding boxes. The other is that we consider the correlation among vehicle parts by GCN rather than directly concatenating part-level features.

Table 5: Comparison of the state-of-the-art vehicle Re-Id methods on the VehicleID dataset.

Settings	Small		Medium		Large	
Methods	R-1	R-5	R-1	R-5	R-1	R-5
DRDL [18]	48.9	73.5	42.8	66.8	38.2	61.6
NuFACT [20]	48.9	69.5	43.6	65.3	38.6	60.7
VAMI [44]	63.1	83.3	52.9	75.1	47.3	70.3
C2F [7]	61.1	81.7	56.2	76.2	51.4	72.2
FDA-Net [24]	-	-	59.8	77.1	55.5	74.7
AAVER [11]	74.7	93.8	68.6	90.0	63.5	85.6
MLSL [1]	74.2	88.4	69.2	81.5	66.6	78.7
OIFE [32]	-	-	-	-	67.0	82.9
PRN [8]	78.4	92.3	75.0	88.3	74.2	86.4
PCRNet (ours)	**86.6**	**98.1**	**82.2**	**96.3**	**80.4**	**94.2**

Table 6: Comparison of the state-of-the-art vehicle Re-Id methods on the VERI-Wild dataset.

Settings	Small		Medium		Large	
Methods	mAP	R-1	mAP	R-1	mAP	R-1
GoogLeNet [37]	24.3	57.2	24.2	53.2	21.5	44.6
DRDL [18]	22.5	57.0	19.3	51.9	14.8	44.6
FDA-Net [24]	35.1	64.0	29.8	57.8	22.8	49.4
MLSL [1]	46.3	86.0	42.4	83.0	36.6	77.5
PCRNet (ours)	**81.2**	**92.5**	**75.3**	**89.6**	**67.1**	**85.0**

Evaluation on VehicleID and VERI-Wild. Table 5 and Table 6 list the results on VehicleID and VERI-Wild, respectively. The results also reflect the trend that the methods which exploit both global and local features achieve better results than early methods that only use global features. Meanwhile, our PCRNet obtains the state-of-the-art performance on VehicleID and VERI-Wild, which shows the effectiveness and generalization of our method.

5.4 Ablation Study and Discussion

In this section, we conduct the ablation study of the components in the PCRNet on the VeRi dataset. We first compare the combinations of different network structures, i.e., the global branch (**G**), the local branch with GCN (**GCN**), and the local branch with self-supervised GCN (**SelfGCN**). We also compare the performance between the proposed part erasing augmentation (**PE**) and the widely used random erasing augmentation (**RE**).

The results of the ablation study are listed in Table 7. First of all, through the comparison of three individual networks, i.e., G, GCN, and SelfGCN, we can find that the global branch significantly outperforms the local branch. This means that the global features learned by deep CNNs are effective for vehicle Re-Id, while only depending on the local features is insufficient. Moreover, by combining the global and local branches, the results are better than the individual branches, which proves that the global appearances and local details are complementary for discriminative representations. Besides, the self-supervised learning strategy makes the GCN more powerful to learn robust local features and brings 2.0% increase of mAP than only using the global branch. Furthermore,

Table 7: The ablation study of PCRNet on the VeRi dataset.

G	GCN	SelfGCN (ours)	RE	PE	mAP	R-1	R-5
✓					75.5	94.5	97.8
	✓				62.5	86.9	95.0
		✓			65.5	85.1	87.6
✓			✓		75.0	94.8	97.9
✓				✓	76.6	95.0	98.2
✓	✓				76.3	94.5	97.6
✓		✓			77.5	95.1	98.1
✓		✓		✓	**78.6**	**95.4**	**98.4**

the comparison between RE and PE demonstrates that the proposed part erasing augmentation can better enhance the generalization of the model for vehicle Re-Id than the general random erasing augmentation. This also reflects that part-level information provides important knowledge for vehicle Re-Id. At last, the overall structure of the PCRNet achieves the state-of-the-art performance on the VeRi dataset, which shows its superiority for vehicle Re-Id.

Albeit the significant effectiveness, since our PCRNet has two separate branches for global and local representations, the computation cost is higher than the single-branch structure. Therefore, in future work, we will try to share part of the CNN layers of the two backbone networks, which can reduce the number of parameters while maintaining the capability of representation learning.

6 CONCLUSION

In this paper, we propose a Parsing-guided Cross-part Reasoning Network, dumbed as PCRNet, which explores the accurate parsed parts to discover the relations among vehicle parts and learn discriminative local features for vehicle re-identification. For the global branch of the PCRNet, the deep CNN learns the overall representation of vehicles with a random part erasing augmentation, which makes the model more robust to occlusion and viewpoint variations. For the local branch, we design a parsing-guided cross-part reasoning module to discover the relations between neighboring parts. This module exploits the GCNs to propagate local information among different parts and learns significant local features for vehicle Re-Id. Moreover, a self-supervised learning strategy is proposed to make the GCNs able to predict features of invisible parts from visible parts. Through extensive experiments, our PCRNet obtains the state-of-the-art performance on three public vehicle Re-Id datasets. In addition, to achieve accurate vehicle parsing, we construct a novel Multi-grained Vehicle Parsing dataset which not only contains large-scale vehicle images annotated with coarse and fine-grained parsing masks, but also reflects diverse variations of environmental factors in real-world surveillance scenes. Beside Re-Id, the MVP dataset can facilitate a wide range of vehicle-related applications, e.g., fine-grained classification, pose estimation, automatic driving, and etc.

7 ACKNOWLEDGEMENT

This work is partially supported by Beijing Academy of Artificial Intelligence (BAAI), Zhejiang Province Nature Science Foundation of China (No. LR17F030006), and National Nature Science Foundation of China (No. 61931008 and No. 61671196).

REFERENCES

[1] Saghir Ahmed Saghir Alfasly, Yongjian Hu, Haoliang Li, Tiancai Liang, Xiaofeng Jin, Beibei Liu, and Qingli Zhao. 2019. Multi-Label-Based Similarity Learning for Vehicle Re-Identification. *IEEE Access* 7 (2019), 162605–162616.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision*. 833–851.

[3] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. 2014. Detect What You Can: Detecting and Representing Objects Using Holistic Models and Body Parts. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1979–1986.

[4] Chuang Gan, Hang Zhao, Peihao Chen, David D. Cox, and Antonio Torralba. 2019. Self-Supervised Moving Vehicle Tracking With Stereo Sound. In *IEEE International Conference on Computer Vision*. 7052–7061.

[5] Qichuan Geng, Feixiang Lu, Xinyu Huang, Sen Wang, Xinjing Cheng, Zhong Zhou, and Ruigang Yang. 2018. Part-level Car Parsing and Reconstruction from Single Street View. *CoRR* abs/1811.10837 (2018).

[6] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6757–6765.

[7] Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. 2018. Learning Coarse-to-Fine Structured Feature Embedding for Vehicle Re-Identification. In *AAAI Conference on Artificial Intelligence*. 6853–6860.

[8] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. 2019. Part-Regularized Near-Duplicate Vehicle Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3997–4005.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[10] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-Aware Pyramid Reconstruction for Alignment-Free Occluded Person Re-Identification. In *IEEE International Conference on Computer Vision*. 8449–8458.

[11] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. 2019. A Dual-Path Model With Adaptive Attention for Vehicle Re-Identification. In *IEEE International Conference on Computer Vision*. 6131–6140.

[12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

[13] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

[14] Jianshu Li, Jian Zhao, Yunpeng Chen, Sujoy Roy, Shuicheng Yan, Jiashi Feng, and Terence Sim. 2018. Multi-Human Parsing Machines. In *ACM International Conference on Multimedia*. 45–53.

[15] Shuangqun Li, Xinchen Liu, Wu Liu, Huadong Ma, and Haitao Zhang. 2016. A discriminative null space based deep learning approach for person re-identification. In *International Conference on Cloud Computing and Intelligence Systems*. 480–484.

[16] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2019. Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 4 (2019), 871–885.

[17] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2197–2206.

[18] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. 2016. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2167–2175.

[19] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2016. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In *European Conference on Computer Vision*. 869–884.

[20] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2018. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Trans. Multimedia* 20, 3 (2018), 645–658.

[21] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. 2018. RAM: A Region-Aware Deep Model for Vehicle Re-Identification. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[22] Yinglu Liu, Hailin Shi, Yue Si, Hao Shen, Xiaobo Wang, and Tao Mei. 2019. A High-Efficiency Framework for Constructing Large-Scale Face Parsing Benchmark. *CoRR* abs/1905.04830 (2019).

[23] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. 2019. Embedding Adversarial Learning for Vehicle Re-Identification. *IEEE Trans. Image Processing* 28, 8 (2019), 3794–3807.

[24] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. 2019. VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3235–3243.

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1487–1495.

[26] Silvio Savarese and Fei-Fei Li. 2007. 3D generic object categorization, localization and pose estimation. In *IEEE International Conference on Computer Vision*. 1–8.

[27] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. 2017. Learning Deep Neural Networks for Vehicle Re-ID with Visual-spatio-Temporal Path Proposals. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 1918–1927.

[28] Brandon M. Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. 2013. Exemplar-Based Face Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3484–3491.

[29] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5693–5703.

[30] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. 2019. PAMTRI: Pose-Aware Multi-Task Learning for Vehicle Re-Identification Using Highly Randomized Synthetic Data. In *IEEE International Conference on Computer Vision*. 211–220.

[31] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David C. Anastasiu, and Jenq-Neng Hwang. 2019. CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 8797–8806.

[32] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. 2017. Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification. In *IEEE International Conference on Computer Vision*. 379–387.

[33] Hongtao Xie, Zhendong Mao, Yongdong Zhang, Han Deng, Chenggang Yan, and Zhineng Chen. 2019. Double-Bit Quantization and Index Hashing for Nearest Neighbor Search. *IEEE Trans. Multimedia* 21, 5 (2019), 1248–1260.

[34] Kota Yamaguchi, M. Hadi Kiapour, and Tamara L. Berg. 2013. Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items. In *IEEE International Conference on Computer Vision*. 3519–3526.

[35] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. 2012. Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3570–3577.

[36] Chenggang Yan, Liang Li, Chunjie Zhang, Bingtao Liu, Yongdong Zhang, and Qionghai Dai. 2019. Cross-Modality Bridging and Knowledge Transferring for Image Understanding. *IEEE Trans. Multimedia* 21, 10 (2019), 2675–2685.

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

[38] Runhao Zeng, Wenbing Huang, Chuang Gan, Mingkui Tan, Yu Rong, Peilin Zhao, and Junzhou Huang. 2019. Graph Convolutional Networks for Temporal Action Localization. In *IEEE International Conference on Computer Vision*. 7093–7102.

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6230–6239.

[40] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2018. Understanding Humans in Crowded Scenes: Deep Nested Adversarial Learning and A New Benchmark for Multi-Human Parsing. In *ACM International Conference on Multimedia*. 792–800.

[41] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *IEEE International Conference on Computer Vision*. 1116–1124.

[42] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. 2017. Person Re-identification in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3346–3355.

[43] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random Erasing Data Augmentation. *CoRR* abs/1708.04896 (2017).

[44] Yi Zhou and Ling Shao. 2018. Viewpoint-Aware Attentive Multi-View Inference for Vehicle Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6489–6498.