Part-Guided Attention Learning for Vehicle Re-Identification*

Xinyu Zhang°, Rufeng Zhang°, Jiewei Cao†, Dong Gong†, Mingyu You°, Chunhua Shen†

†The University of Adelaide, Australia °Tongji University, China

Abstract

Vehicle re-identification (Re-ID) often requires one to recognize the fine-grained visual differences between vehicles. Besides the holistic appearance of vehicles which is easily affected by the viewpoint variation and distortion, vehicle parts also provide crucial cues to differentiate near-identical vehicles. Motivated by these observations, we introduce a *Part-Guided Attention Network* (PGAN) to pinpoint the prominent part regions and effectively combine the global and part information for discriminative feature learning. PGAN first detects the locations of different part components and salient regions regardless of the vehicle identity, which serve as the *bottom-up attention* to narrow down the possible searching regions. To estimate the importance of detected parts, we propose a *Part Attention Module* (PAM) to adaptively locate the most discriminative regions with high-attention weights and suppress the distraction of irrelevant parts with relatively low weights. The PAM is guided by the Re-ID loss and therefore provides *top-down attention* that enables attention to be calculated at the level of car parts and other salient regions. Finally, we aggregate the global appearance and part features to improve the feature performance further. The PGAN combines part-guided bottom-up and top-down attention, global and part visual features in an end-to-end framework. Extensive experiments demonstrate that the proposed method achieves new state-of-the-art vehicle Re-ID performance on four large-scale benchmark datasets.

1 Introduction

Vehicle re-identification (Re-ID) aims to verify whether or not two vehicle images captured by different surveillance cameras belong to the same identity. With the growth of road traffic, it plays an increasingly important role in urban security systems and intelligent transportation [1, 2, 3, 4, 5, 6, 7, 8].

Different levels of granularity of visual attention are required under various Re-ID scenarios. In the case of comparing vehicles of different car models, we can easily distinguish their identities by examining the overall appearances, such as car types and headlights [3]. However, most production vehicles can exhibit near-identical appearances since they may be mass-produced by the same manufacturer. When two vehicles with the same car model are presented, more fine-grained details (*e.g.*, annual service signs, customize paintings, and personal decorations) are required for comparison, as shown in Figure 1 (a). Therefore, the key challenge of vehicle Re-ID lies in how to recognize the subtle visual differences between vehicles and locate the prominent parts that characterize their identities.

Most existing works focus on learning global appearance features with various vehicle attributes, including model type [9, 6, 10], license plate [9], spatial-temporal information [11, 12], orientation [5, 13, 14, 15], *etc*. The main disadvantage of global features is the lack of capability to capture more fine-grained visual differences, which is crucial in vehicle Re-ID. Also, they are easily degraded by the viewpoint variation, distortion, occlusion, motion blur and illumination, especially in the unconstrained real-world environment. Therefore, recent works tend to explore car parts [16, 17, 18] to learn the local information. However, these methods mainly focus on the localization of the spatial regions without considering how these regions are subject to attention with different degree.

To address the problems above, we propose a novel *Part-Guided Attention Network* (PGAN) to detect the prominent part regions of a vehicle at a sufficiently fine-grained level and combine these part characteristics with the holistic appearance for discriminative feature learning. To better capture the details of different vehicle components, the proposed PGAN first detects various parts and salient regions regardless of the vehicle identity, *e.g.*, car lights, logos, and annual service signs, as shown in Figure 1 (b). These detected parts serve as candidate regions for comparison, which help narrow down the possible searching regions for network learning. The part extraction module is instantiated by

*This work was done when X. Zhang was visiting The University of Adelaide. Correspondence should be addressed to C. Shen.

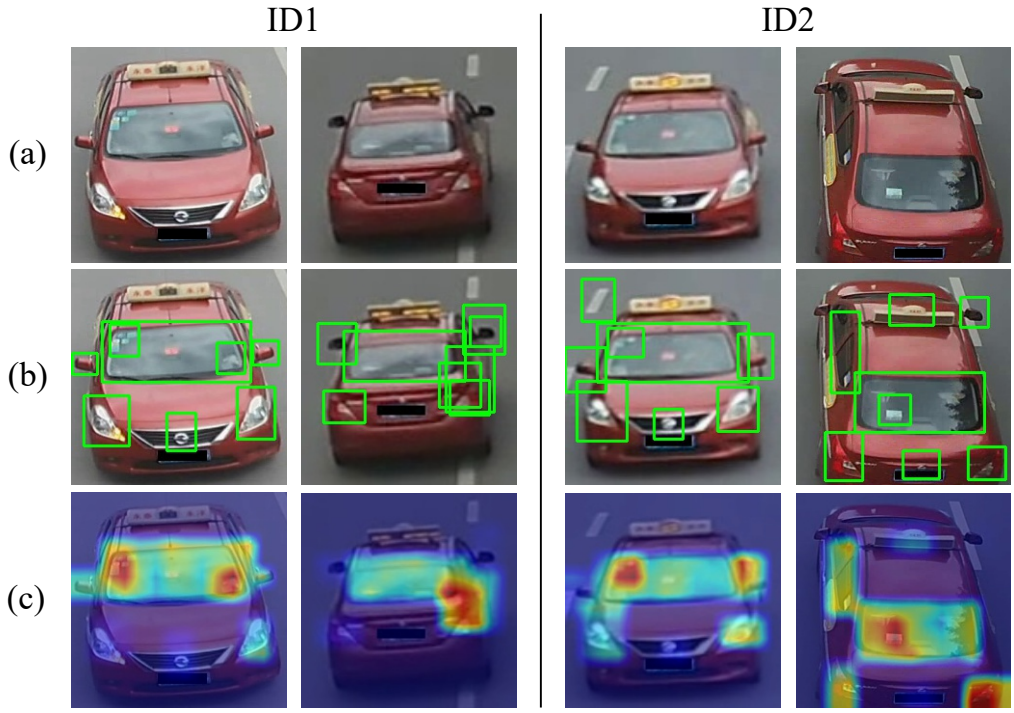

Figure 1: Illustration of the part-guided attention. (a) The rear and front views of two different vehicles with the same car model. (b) The detected candidate part regions from the part extraction module. (c) The heatmaps of part features from the part attention module. The prominent part regions like annual signs are highlighted, while the wrong candidates and insignificant parts like background and back mirror are suppressed.

an object detection network pre-trained on the vehicle attributes datasets [3, 18]. We refer this kind of attention as *bottom-up attention* since it is not driven by the Re-ID task, but instead is determined by the region saliency and vehicle attributes.

The next important step of PGAN is to select the most prominent part regions and assign appropriate attention scores to them. Here, we introduce a *Part Attention Module* (PAM) to provide top-down attention guided by the Re-ID loss. PAM adaptively locates the discriminative regions with high-attention weights and suppresses the distraction of irrelevant parts with relatively low weights, as shown in Figure 1 (c). Compared to the existing grid attention or evenly decomposed part attention [19, 20, 17, 21], our PGAN is able to provide more fine-grained attention which is calculated at the level of car parts and subtle areas. The success of combining bottom-up and top-down attention has also been witness in image captioning and visual question answering, where the fine-grained analysis and image understanding are required [22]. Different from these works, our attention mechanism is tailored to handle the car parts and visual attributes that are non-negligible for vehicle Re-ID. Finally, we aggregate the vehicle's holistic appearance and part characteristics with a features aggregation module to improve the feature performance further.

To summarize, our main contributions are as follows:

- We design a novel Part-Guided Attention Network (PGAN) to capture the fine-grained part details and learn discriminative features for vehicle Re-ID. The PGAN combines part-guided bottom-up and top-down attention, global and part features in an end-to-end framework.
- We propose a Part Attention Module (PAM) to pay more attention to the prominent parts adaptively, and reduce the distraction of wrongly detected or irrelevant parts.
- Extensive experiments on four challenging benchmark datasets demonstrate that our proposed method achieves new state-of-the-art vehicle Re-ID performance.

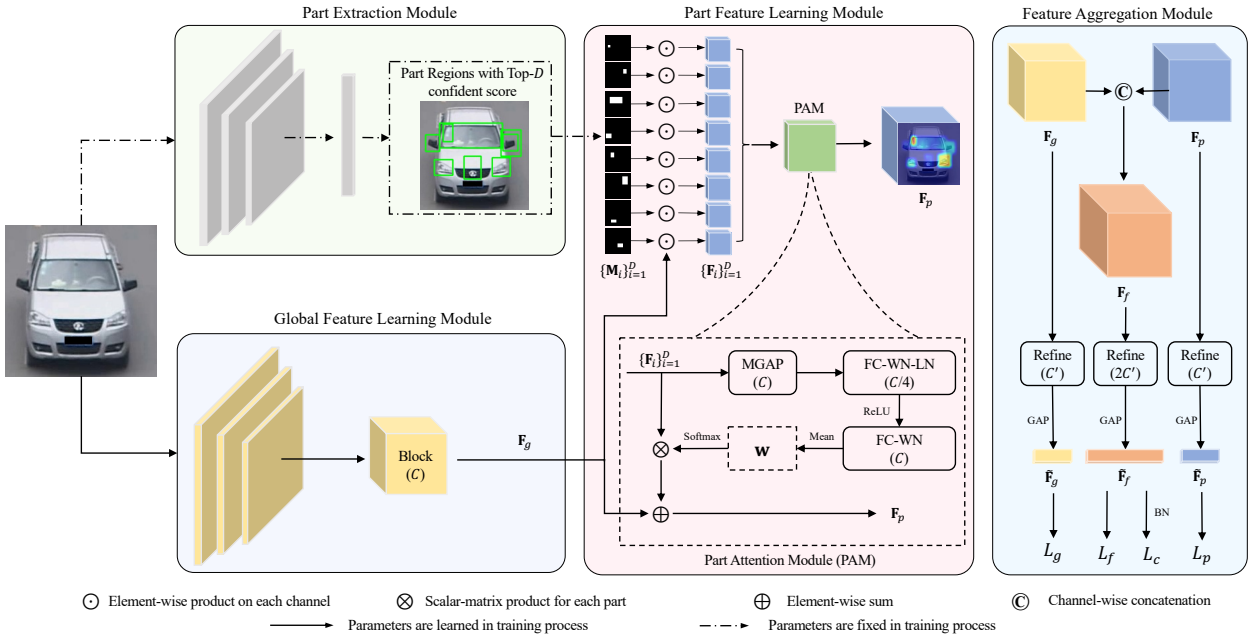

Figure 2: Part-Guided Attention Network (PGAN) pipeline. The model consists of four modules: Part Extraction Module, Global Feature Learning Module, Part Feature Learning Module and Feature Aggregation Module. The input vehicle image is first processed to obtain the global feature $\mathbf{F}_g$ and the part masks $\{\mathbf{M}_i\}_{i=1}^{D}$ of Top-D candidate parts. The part mask features $\{\mathbf{F}_i\}_{i=1}^{D}$ is then obtained via Eq. (2), after which $\{\mathbf{F}_i\}_{i=1}^{D}$ is fed into a Part Attention Module (PAM) to obtain the part-guided feature $\mathbf{F}_p$. PAM is a compact network, learning a soft attention weight $\mathbf{w} \in \mathbb{R}^D$, which is composed of a mask-guided average pooling (MGAP) layer and some linear and non-liner layers. Subsequently, the fusion feature $\mathbf{F}_f$ is obtained by concatenating $\mathbf{F}_g$ and $\mathbf{F}_p$. After the refinement and global average pooling (GAP) operation, $\widetilde{\mathbf{F}}_g$, $\widetilde{\mathbf{F}}_p$ and $\widetilde{\mathbf{F}}_f$ are all used for optimization. L_f, L_g and L_p are triplet loss functions while L_c is a softmax cross-entropy loss. Here, FC, WN, LN and BN represent fully-connected layer, weight normalization, layer normalization and batch normalization respectively. Mean denotes a channel-wise mean operation. C and C' are channel dimension before and after refine operation.

2 Related Work

2.1 Global Feature-based Methods

Feature Representation Vehicle Re-ID aims at learning discriminative feature representation to deal with significant appearance changes for different vehicles. Public large-scale datasets [4, 6, 9, 8, 23, 8, 24, 3] are widely collected with annotated labels and abundant attributes under unrestricted conditions. These datasets face huge challenges on occlusion, illumination, low resolution and various views. One way to deal with these datasets uses deep features [5, 9, 25, 24, 4] instead of hand-crafted features to describe vehicle images. To learn more robust features, some methods [9, 6, 26, 11, 12, 10] try to explore details of vehicles using additional attributes, such as model type, color, spatial-temporal information, *etc*. Moreover, works of [15, 13] propose to use synthetic multi-view vehicle images from a generative adversarial network (GAN) to alleviate cross-view influences among vehicles. In [14, 5] authors also implement view-invariant inferences effectively by learning a viewpoint-aware representation. Although great progress has been obtained by these methods, there is a huge drop when encountering invisible variance of different vehicles as well as large diversity in same vehicle identity.

Metric Learning To alleviate the above limitation, deep metric learning methods [27, 28, 29, 30, 29] use powerful distance metric expression to pull vehicle images in the same identity closer while pushing dissimilar vehicle images further away. The core idea of these methods is to utilize the matching relationship between image pairs or triplets as much as possible. Whereas, sampling strategies in deep metric learning lead to suboptimal results and also lack of abilities to recognize more meaningful unobtrusive details.

2.2 Part Feature-based Methods

Beyond learning global distinguishable features, a series of part-based learning methods explicitly exploit the discriminative information from multi-part locations of vehicles. [21, 19, 17, 20] take great efforts on separating feature maps into multiple even partitions to extract specific feature representation of respective regions. Another line of part-based methods [31, 32, 33] bring informative key-points to put more attention on effective localized features. Besides, [18, 16] denote to design part-fused networks using ROI features of each part on vehicles from a pre-trained detection model to extract discriminative features. Nevertheless, these methods merely pay attention to exploring the part locations while ignoring the consideration of the important degree of the different part regions.

3 Methodology

We firstly define each vehicle image as x and the unique corresponding identity label as y. Given a training set X^t=$\{(x_n^t, y_n^t)\}_{n=1}^{N^t}$, the main goal of the vehicle Re-ID is to learn a feature embedding function $\phi(x^t;\ \theta)$ for measuring the vehicle similarity under certain metrics, where θ denotes the parameters of $\phi(\cdot)$. It is important to learn a ϕ with good generalization on unseen testing images. During testing, given a query vehicle image x^q, we can find vehicles with the same identity from a gallery set X^g = $\{(x_n^g, y_n^g)\}_{n=1}^{N_g}$ by comparing the similarity between $\phi(x^q;\theta)$ and each $\phi(x_n^g;\theta),\ \forall x_n^g$.

In this section, we present the proposed Part-Guided Attention Network (PGAN) in detail. The overall framework is illustrated in Figure 2, which consists of four main components: *Part Extraction Module*, *Global Feature Learning Module*, *Part Feature Learning Module* and *Feature Aggregation Module*. We first generate the part masks of vehicles in the part extraction module, which are then applied on the global feature map to obtain the mask-guided part feature. After that, we learn the attention scores of different parts to enhance the part feature via increasing the weights of discriminative parts as well as decreasing that of less informative parts. Subsequently, the three refined features, *i.e.*, global, part, and fusion features are all used for model optimization.

3.1 Global Feature Learning Module

For a vehicle image x, before obtaining the part features, we first extract a global feature map $\mathbf{F}_g \in \mathbb{R}^{H\times W\times C}$ with a standard convolutional neural network, as shown in Figure 2. Most previous methods [34, 35] directly feed $\mathbf{F}_g$ into a global average pooling (GAP) layer to obtain the embedding feature that mainly considers the global information, which is studied as a baseline model in our experiments.

However, maintaining the spatial structure of feature map helps describe the subtle visual differences, which is crucial for distinguishing two near-identical vehicles. Therefore, we directly apply $\mathbf{F}_g$ as one of the inputs for the following part learning process and the final optimization.

3.2 Part Extraction Module

We extract the part regions using a pre-trained SSD detector specially trained on vehicle components [18], which is a fast one-stage detector. In the part extraction, 16 vehicle attributes (*e.g.*, annual signs, car lights, logos, and entry license) are considered. Details are left in the supplementary material. Once detected, we only use the confidence scores to select part regions and ignore the label information of each part. It is reasonable since not all attributes are available in each vehicle due to the view variation.

Instead of naively selecting relevant part regions by thresholding the confidence score, we select the most confident top-D proposals as the candidate vehicle parts. The main reasons are twofold: 1) some crucial yet less confident bounding boxes, like annual service signs, play a crucial role in distinguishing different vehicle images; 2) part number is fixed, which is easy to learn the attention model in the following stage. Note that we want to ensure a high recall rate to avoid missing relevant parts. The irrelevant parts are filtered out from the subsequent attention learning. Figure 1 illustrates some vehicle samples with the selected candidate parts. More visualizations are in supplementary material.

We use the index $i \in \{1, 2, ..., D\}$ to indicate each of the selected top-D part regions. The spatial area covered by each part is denoted as A_i. For each candidate part region i, we obtain a binary mask matrix $\mathbf{M}_i \in \{0, 1\}^{H\times W}$ by assigning 1 to the elements inside the part region and 0 to the rest:

$$\mathbf{M}_i(\text{pix}) = \begin{cases} 1, & \text{if pix} \in A_i \\ 0, & \text{if pix} \notin A_i \end{cases}, \forall i, \tag{1}$$

where pix indicates a pixel location of $\mathbf{M}_i$. Note that the size of each $\mathbf{M}_i$ is the same as a single channel of $\mathbf{F}_i$. If the neural network or the size of input image changes, the corresponding part locations on $\mathbf{M}_i$ will be changed accordingly. During processing, we force all A_i in the range of $H \times W$ to ensure all part regions are located in the image areas.

After obtaining global feature $\mathbf{F}_g$ and part masks $\{\mathbf{M}_i\}_{i=1}^{D}$, we project the part masks on the feature map $\mathbf{F}_g$ to generate a set of mask-based part feature representations $\{\mathbf{F}_i\}_{i=1}^{D}$, which will be taken as the input of the following part feature learning module. For each part region i, we can obtain $\mathbf{F}_i$ as:

$$\mathbf{F}_i = \mathbf{M}_i \odot \mathbf{F}_g, \ \ \forall i \in \{1, 2, ..., D\}, \tag{2}$$

where $\odot$ denotes the element-wise product operation on each channel of $\mathbf{F_g}$. $\mathbf{F}_i$ is the mask-based part feature map of the i-th part region. Note that all $\mathbf{F}_i \in \mathbb{R}^{H \times W \times C}$. In each $\mathbf{F}_i$, only the elements in the regions of i-th part are activated.

We learn an attention module on the part regions in the following section. Unlike the traditional grid attention method that processes a set of uniform grids, our attention model can focus on the prominent parts by only activating the selected parts. The irrelevant parts can thus be ignored directly. And the context correlation in a same part can be integrally considered, alleviating missing of essential features. Moreover, this part extraction process can be considered as bottom-up attention [22].

3.3 Part Feature Learning Module

Part feature learning module is to produce a weight map across the mask-based part feature maps $\{\mathbf{F}_i\}$. In this way, the network can take more attention to specific part regions. A recent work [16] simply uses the part mask features as the input of a subsequent feature learning process. In [36], all the elements inside the detected parts are assigned a fixed larger weight than those in other regions. However, all these methods treat different part regions equally and thus more prominent parts cannot be further highlighted . On the other hand, some detected parts might not be informative for some specific cases, such as wrongly detected background or windshield with no specific information, which tends to influence the results. Consequently, we propose a *Part Attention Module* (PAM) to adaptively learn the importance of each part so as to take more attention to the most discriminating part regions and suppress the less informative parts. Since this attention signal is supervised by the specific Re-ID task, it can be considered as part-based top-down attention.

Part Attention Module (PAM) Our PAM is designed to obtain a part-guided feature representation $\mathbf{F}_p \in \mathbb{R}^{H \times W \times C}$ relying on a soft attention mechanism on each part. When we have a soft attention weight vector $\mathbf{w} \in \mathbb{R}^{D}$ to indicate the importance of each part region, we can obtain the part-guided feature representation $\mathbf{F}_p$ as:

$$\mathbf{F}_p = \sum_{i=1}^{D} w_i \mathbf{F}_i + \mathbf{F}_g, \tag{3}$$

where $w_i \in [0, 1]$ denotes the i-th element of the attention weight $\mathbf{w}$, which represents a learned weight of i-th part obtained via Eq. (4). $\mathbf{w}$ is normalized with sum as 1 so that the relative importance between different parts is obvious. $\mathbf{F}_g$ is added to augment the capability of part regions.

We learn a compact model to predict the attention weights $\mathbf{w}$ for measuring the different importance of each selected part, as shown in Figure 2. Specifically, we first use a mask-guided global average pooling operation on each $\mathbf{F}_i$ and then learn a mapping function with a softmax layer to obtain $\mathbf{w}$. Each element w_i can be predicted by:

$$w_i = \frac{\exp(\ \psi(\mathrm{mgap}(\mathbf{F}_i, \mathbf{M}_i), \theta_\psi)\ )}{\sum_{i=1}^{D} \exp(\ \psi(\mathrm{mgap}(\mathbf{F}_i, \mathbf{M}_i), \theta_\psi)\ )}, \tag{4}$$

where $\psi(\cdot)$ denotes a learnable function that is able to highlight the most important part regions with high values (as shown in Figure 2). θ_ψ is the parameter of $\psi(\cdot)$, and $\mathrm{mgap}(\cdot)$ denotes the mask-guided global average pooling (MGAP) discussed in the following.

Before feeding $\mathbf{F}_i$ into ψ, we average each channel of $\mathbf{F}_i$ as a scalar via the $\mathrm{mgap}(\cdot)$ operator. Note that, in each $\mathbf{F}_i$, only the elements in the part region i are activated and most of the elements in $\mathbf{F}_i$ are zero. Instead of performing the standard global average pooling (GAP), we restrict the average pooling in the areas indicated by the mask $\mathbf{M}_i$ via the MGAP operator. In detail, for each channel of $\mathbf{F}_i$, after summing the nonzero elements, the MGAP operator devides the sum value with the number of elements (*i.e.*$\|\mathbf{M}_i\|_1 < H \times W$), instead of the number of total elements of the feature channel (*i.e.* $H \times W$) in the GAP.

3.4 Feature Aggregation Module

Since global and part-based features provide complementary information, we concatenate the global feature $\mathbf{F}_g$ and part-guided feature $\mathbf{F}_p$ together, which is then denoted as fusion feature $\mathbf{F}_f \in \mathbb{R}^{H \times W \times 2C}$. Furthermore, we adopt a *Refine* operation on $\mathbf{F}_f$ to reduce the dimension of feature representation to speed up the training process. The *Refine* operation is composed of a SE Block [37] and a Residual Block [38]. Finally, after a GAP layer, the refined fusion feature $\widetilde{\mathbf{F}}_f \in \mathbb{R}^{2C'}$ is obtained as the feature representation.

3.5 Model Training

In training process, we use Softmax cross-entropy loss and Triplet loss [34] as a joint optimization for $\widetilde{\mathbf{F}}_f$, which are denoted as L_c and L_f. Note that following [35], an additional batch normalization(BN) operation is adopted on $\widetilde{\mathbf{F}}_f$ for Softmax cross-entropy loss. The normalized fusion feature $\widetilde{\mathbf{F}}_f$ is used as the feature representation for evaluation in our work. In order to make full use of global and part information separately, we also optimize the refined global feature $\widetilde{\mathbf{F}}_g \in \mathbb{R}^{C'}$ and the refined part-guided feature $\widetilde{\mathbf{F}}_p \in \mathbb{R}^{C'}$ with Triplet loss function [34]. We define these two loss functions as L_g and L_p, respectively. The total loss function can be formulated as:

$$L = \lambda L_c + L_f + L_g + L_p, \tag{5}$$

where λ is the loss weight to trade off the influence of two types of loss functions. Experiments show that joint optimization could improve the ability of feature representation.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our PGAN method on four public large-scale Vehicle Re-ID benchmark datasets.

VeRi-776 [9] is a challenging benchmark in vehicle Re-ID task that contains about $50,000$ images of 776 vehicle identities across 20 cameras. Each vehicle is from 2-18 cameras with various viewpoints, illuminations and occlusions. All datasets are split into a training set with $37,778$ images of 576 vehicles and a testing set with $11,579$ images with 200 vehicles.

VehicleID [23] is a widely-used vehicle Re-ID dataset which contains vehicle images captured in the daytime by multiple cameras. There are total of $221,763$ images with $26,267$ vehicles, where each vehicle has either front or rearview. The training set contains $110,178$ images of $13,134$ vehicles while the testing set comprises $111,585$ images of $13,133$ vehicles. There are three test subsets with different sizes, *i.e.*, $7,332$ images of 800 IDs in small test set, $12,995$ images of $1,600$ vehicles in medium test set and $20,038$ images of $2,400$ vehicles in large test set.

VRIC [24] is a realistic vehicle Re-ID benchmark with unconstrained variations of images in resolution, motion blur, illumination, occlusion, and viewpoint. It contains $60,430$ images of $5,622$ vehicle identities captured from 60 different traffic cameras during both daytime and nighttime. The training set has $54,808$ images of $2,811$ vehicles, while the rest is used for testing with $5,622$ images of another $2,811$ vehicle IDs.

VERI-Wild [8] is recently released with $416,314$ vehicle images of $40,671$ IDs captured by 174 cameras. The training set consists of $30,671$ IDs with $277,797$ images. Similar as VehicleID, the small test subset consists of $3,000$ IDs with $41,816$ images and the medium/large subset consists of $5,000/10,000$ IDs with $69,389/138,517$ images.

Evaluation Metrics. To measure the performance for vehicle Re-ID, we utilize the Cumulated Matching Characteristics (CMC) and the mean Average Precision (mAP) as evaluation criterions. The CMC calculates the cumulative percentage of correct matches appearing before the top-K candidates. We report Top-1 and Top-5 scores to represent the CMC criterion. Given a query image, Average Precision (AP) is the area under the Precision-Recall curve while mAP is the mean value of AP across all query images. The mAP criterion reflects both precision and recall, which provides a more convincing evaluation on Re-ID task.

4.2 Implementation Details

Part Extraction. We use the same SSD model as [18] to extract part regions. The model is fixed at the training process. For each image, we extract Top-D part regions according to confident scores. In this paper, we set $D = 8$.

Method	mAP	Top-1	Top-5
Baseline	75.7	95.2	98.2
$\widetilde{\mathbf{F}}_f$	77.7	95.9	**98.5**
$\widetilde{\mathbf{F}}_f$ + $\widetilde{\mathbf{F}}_g$	78.0	95.1	97.7
$\widetilde{\mathbf{F}}_f$ + $\widetilde{\mathbf{F}}_p$	78.5	95.8	98.3
$\widetilde{\mathbf{F}}_f$ + $\widetilde{\mathbf{F}}_g$ + $\widetilde{\mathbf{F}}_p$ (PGAN)	**79.3**	**96.5**	98.3

Table 1: Performance comparison on the effectiveness of feature aggregation with different features of PGAN on VeRi-776.

Method	Dimension	mAP	Top-1	Top-5
Grid Attention	256	76.1	95.3	97.7
PGAN w/o PAM	256	77.9	95.6	**98.4**
PGAN	256	78.6	95.4	98.0
Grid Attention	512	77.0	95.8	98.0
PGAN w/o PAM	512	78.0	95.5	98.2
PGAN	512	**79.3**	**96.5**	98.3

Table 2: Performance comparison on different attention methods, *i.e.*, grid attention, PGAN without Part Attention Module (PAM) and our PGAN on VeRi-776.

Vehicle Re-ID. We adopt ResNet50 [38] without the last classification layer as backbone model in global feature learning module, which is pre-trained on ImageNet [39]. The model modification follows [35], and a refined model is added for a fair comparison, which is called *baseline* model in our work.

All input images are resized to 224×224 while only random horizontal flipping and random erasing [40] with a probability of 0.5 are applied for data augmentation. We use Adam optimizer [41] with a momentum of 0.9 and a weight decay 5×10^{-4}. For all experiments without other specification, we set the batch size to 64 with 16 vehicle IDs randomly selected. The learning rate starts from 1.75×10^{-4} and is multiplied by 0.5 every 20 epochs. The total number of epochs is 130.

4.3 Ablation Study

We conduct extensive experiments on VeRi-776 to thoroughly analyze the effectiveness of our PGAN method.

4.3.1 Effectiveness of Feature Aggregation

To validate the necessity of different features in our proposed PGAN, we first design an ablation experiment analyzing the effectiveness of global, part, and fusion feature. We fix the feature dimension of $\widetilde{\mathbf{F}}_f$ to 512. For a fair comparison, we also set the feature dimension to 512 in a baseline model. As reported in Table 1, we can observe that only using $\widetilde{\mathbf{F}}_f$ for optimization can improve the performance by 2% on mAP comparing with baseline model, which confirms that PAM can provide important part information that is better for model optimization. After adding $\widetilde{\mathbf{F}}_g$ and $\widetilde{\mathbf{F}}_p$ separately, mAP can improve by about 1%. It shows that combining with global and part feature can provide more useful information. Furthermore, with the joint optimization with all these features, the result improves to 79.3% and 96.5% on mAP and Top-1, which outperforms baseline model by 3.6% and 1.3%.

4.3.2 Analysis of Different Attention Method

We first implement traditional grid attention by removing part extraction module, *i.e.*, PAM is directly used on each grid of $\mathbf{F}_g \in \mathbb{R}^{H \times W \times C}$. As shown in Table 2, grid attention can only achieve 77.0% mAP and 95.8% Top-1 accuracy when feature dimension is 512, showing that part guidance is crucial for filtering invalid information like background. Moreover, we also use the identical weight for each part region by removing PAM. It can be seen as a bottom-up attention with the part guidance from a detection model. From Table 2, we can find 0.7% and 1.3% mAP decrease

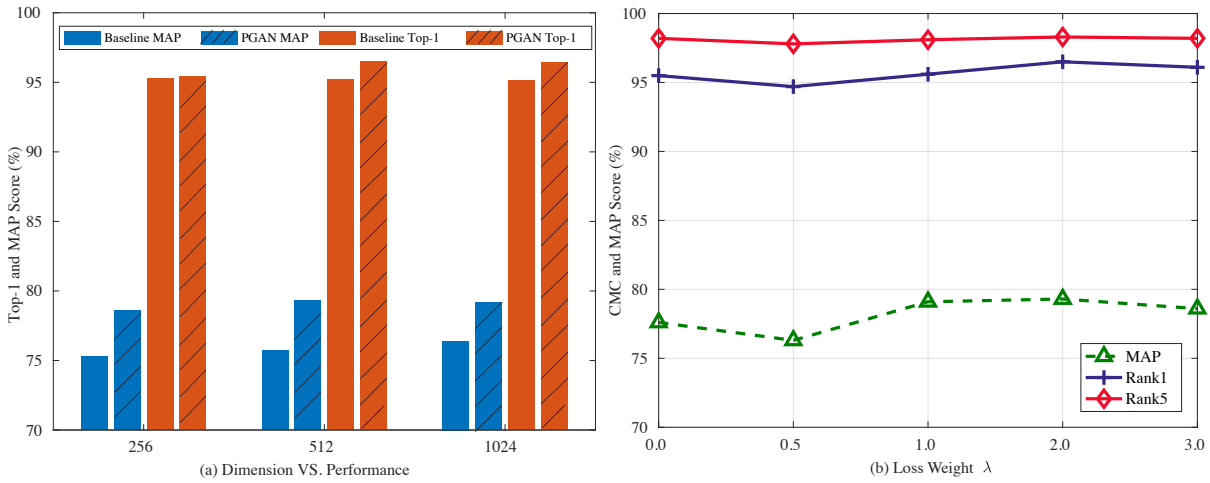

Figure 3: Parameter analysis of PGAN on VeRi-776. (a) Feature dimension and (b) Loss weight λ *vs.* Re-ID accuracy.

when feature dimension is 512 and 256 without PAM. It proves that PAM is beneficial for focusing on prominent parts as well as suppressing the impact of some wrongly detected or useless regions. We exactly note that our PGAN w/o PAM is still better than grid attention by 1.0% mAP, which also proves the important role of the part-guided bottom-up attention. In Figure 4, we visualize one vehicle sample with Top-5 retrieval vehicles and the corresponding heatmaps of the part feature $\mathbf{F}_p$ generated by grid attention and our PGAN respectively. It is clear that our PGAN pays more attention on discriminative part regions. Some wrong detected parts and useless parts like background can be suppressed or ignored. More visualization can be found in the supplementary material.

4.3.3 Parameter Analysis of PGAN

First, we evaluate the effectiveness of different feature dimension. We use the dimension $2C'$ of fusion feature $\widetilde{\mathbf{F}}_f$ on VeRi-776 as the variable. As shown in Figure 3 (a), our PAM module has great improvement compared with the baseline model whatever the dimension is. Note that the improvement is not obtained by increasing the feature dimension. For example, our PGAN with 256 dimension surpasses the baseline model with 512 dimension by a large margin.

Moreover, we conduct experiments to evaluate the impact of loss weight λ in Eq. (5). From Figure 3 (b), we observe that the best result is obtained when λ is set to 2. Also, the value of λ has a limited impact on the performance, and our PGAN is still better than the baseline by a large margin.

4.4 Comparison with State-of-the-art Methods

Finally, we compare our PGAN against other state-of-the-art vehicle Re-ID methods, shown in Table 3. All reported results of our method are based on 512 dimension.

For VeRi-776, we strictly follow the cross-camera-search evaluation protocol as [9]. From Table 3, it is clear that our PGAN outperforms all the existing method for a large margin. For instance, the performance of PGAN is better than the state-of-the-art method, *i.e.*Part-Regular [16], for 5% mAP and 2.3% Top-1 respectively.

For VehicleID, we only report the result of the large test subset on Top-1 and Top-5. Our method surpasses all the method except RNN-HA [10] at Top-1. Notice that RNN-HA uses the additional supervision of the vehicle model and the input of image size is 672×672 (9 times bigger than us). However, as reported in [10], the performance of RNN-HA is extremely dropped by a large margin on VeRi-776 when the image size is set to 224×224, which is lower than our PGAN for about 22% at Top-1.

VRIC and VERI-Wild datasets are newly released large vehicle datasets with more unconstrained variations in resolutions, illuminations, occlusion, and viewpoints, *etc*. There are only a few methods that have reported the results. As for VERI-Wild, we report the result of the large test subset. Table 3 shows that our proposed PGAN achieves satisfactory performance with 78.0% on VRIC and 93.8% on VERI-Wild at Top-1. Compared with the baseline model, our PGAN is more robust under various environments.

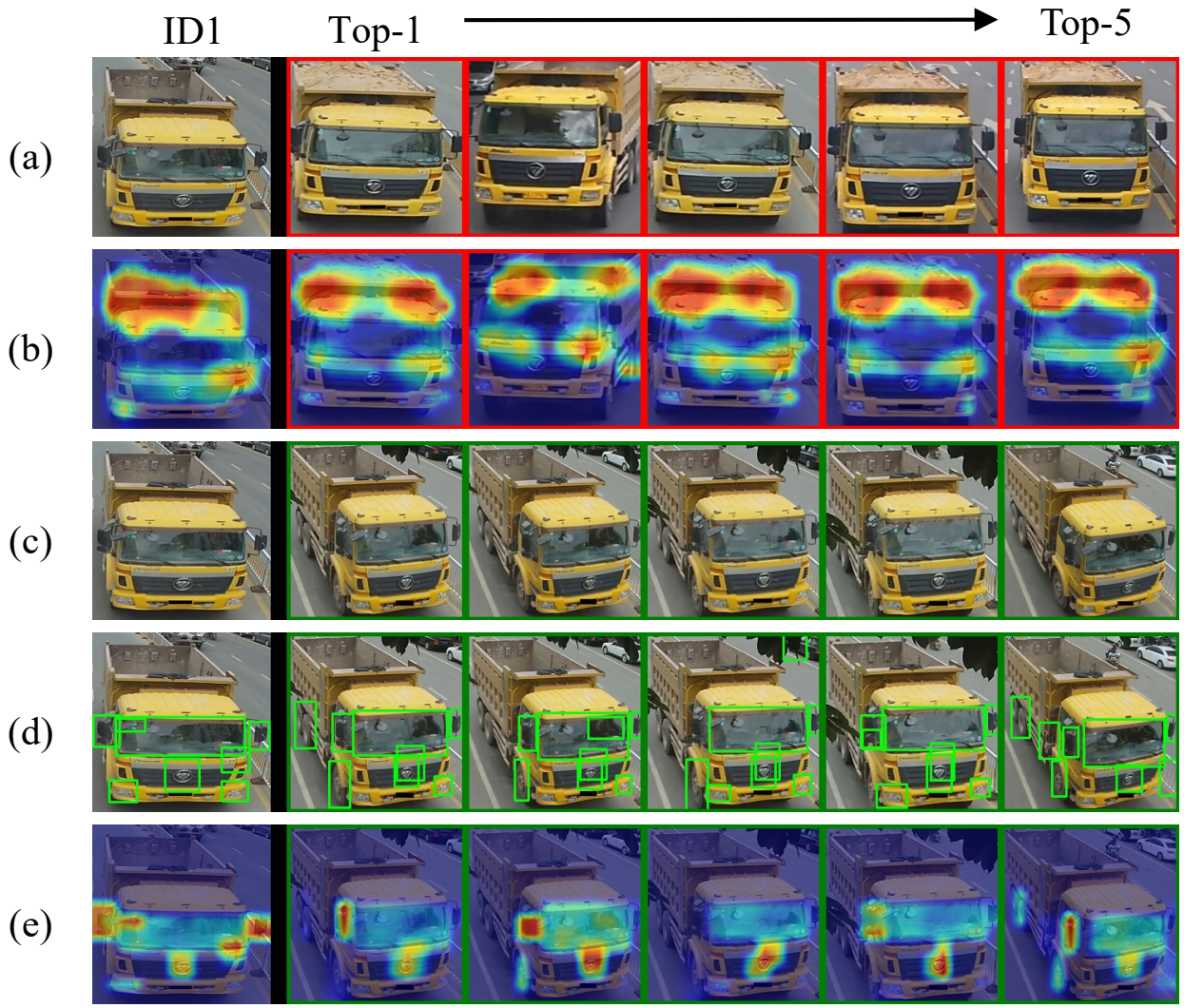

Figure 4: Illustration of visualized comparison between traditional grid attention and our PGAN. For a query image, we draw: (a) Top-5 retrieval results and (b) the corresponding heatmaps of $\mathbf{F}_p$ from PAM in grid attention; (c) Top-5 retrieval results, (d) the detected candidate part regions and (e) the corresponding heatmaps of $\mathbf{F}_p$ from PAM in PGAN. The correct and false matched vehicle images are enclosed in green and red rectangles respectively. It shows that our PGAN can put more attention on the most prominent part regions, such as back mirrors, windshield stickers and car brands. However, the grid attention mainly focuses on some insignificant regions like the car roof, resulting in the attention distracting. (Best viewed in color)

We also report the results of traditional grid attention and PGAN without PAM on each dataset. Experiments show that our method achieves comparable results in all datasets. Especially, PGAN surpasses grid attention by 1.3% at Top-1 accuracy on VRIC, which shows better precision than grid attention. All results prove that our PGAN is able to retrieval more reliably matched vehicle images.

5 Conclusion

In this paper, we have presented a novel Part-Guided Attention Network (PGAN) for vehicle Re-ID. First, we extract part regions of each vehicle image from an object detection model. These part regions provide a range of candidate searching area for the network learning, which is regarded as a bottom-up attention process. Then we use the proposed

Method	VeRi-776		VehicleID		VRIC		VERI-Wild	
	mAP	Top-1	Top-1	Top-5	Top-1	Top-5	mAP	Top-1
FACT+Plate-SNN+STR [9]	27.8	61.4	-	-	-	-	-	-
Siamese-CNN+Path-LSTM [12]	58.3	83.5	-	-	30.6	57.3	-	-
OIFE [5]	51.4	92.4	67.0	82.9	24.6	51.0	-	-
PROVID [11]	53.4	81.6	-	-	-	-	-	-
VAMI [14]	50.1	77.0	47.3	70.3	-	-	-	-
MSVR [24]	49.3	88.6	63.0	73.1	46.6	65.6	-	-
RNN-HA [10]	56.8	74.8	**81.1**	87.4	-	-	-	-
RAM [21]	61.5	88.6	67.7	84.5	-	-	-	-
AAVER [31]	66.4	90.2	63.5	85.6	-	-	-	-
Part-Regular [16]	74.3	94.3	74.2	86.4	-	-	-	-
FDA-Net [8]	55.5	84.3	55.5	74.7	-	-	22.8	49.4
Baseline [35]*	75.7	95.2	77.5	91.0	76.1	93.0	72.9	92.9
Grid Attention	77.0	95.8	77.1	91.4	76.7	93.6	73.6	93.9
PGAN w/o PAM	78.0	95.5	77.6	91.8	77.4	93.1	73.6	93.8
PGAN	**79.3**	**96.5**	77.8	**92.1**	**78.0**	**93.2**	**74.1**	**93.8**

Table 3: Comparisons with state-of-the-art Re-ID methods on VeRi-776, VehicleID, VRIC and VERI-Wild (in %). In each column, the first and second highest results are highlighted by **red** and blue respectively (only PGAN and published methods). The results of Siamese-CNN+Path-LSTM [12] and OIFE [5] on VRIC is reported by MSVR [24]. $*$ represents the baseline model in our paper which is implemented using [35].

part attention module (PAM) to discover the prominent part regions by learning a soft attention weight for each candidate part, which is a top-down attention process. In this way, the most discriminative parts are highlighted with high-attention weights, while the opposite effects of invalid or useless parts are suppressed with relatively low weights. Furthermore, with the joint optimization with the holistic feature and the part-guided feature, the Re-ID performance can be further improved. Extensive experiments are conducted to show the effectiveness of our PGAN. And our PGAN outperforms other state-of-the-art methods by a large margin. In the future, we plan to extend the proposed method to the multi-task learning, *i.e.*, object detection and Re-ID, for simultaneously improving the performance of these two tasks.

A Appendix

A.1 The Attributes of Vehicle Part Regions

The work [18] carefully labelled 21 attributes of vehicles, in which only 16 attributes are adopted in our work. Since the attributes of vehicle style, *i.e.*, "car", "trunk", "tricycle", "train" and "bus", represent the whole vehicle image, they can be recognized as the global information in the area of vehicle Re-ID task. The remaining attributes are shown as Table 4.

Note that we do not use the attribute labels once the detection process is finished in our work since most of vehicles contain only few vehicle parts due to the multi-view variation.

annual service signs	back mirror	car light	carrier
car topwindow	entry license	hanging	lay ornament
light cover	logo	newer sign	tissuebox
plate	safe belt	wheel	wind-shield glass

Table 4: Name of vehicle attributes.

A.2 Analysis of the Number of Part Regions D.

In addition, we analyse how the number of part regions D in the part extraction module affects the Re-ID results. We test the performance with $D = \{0, 4, 6, 8, 10\}$ of our PGAN on VeRI-776. The feature dimension is fixed to 512.

As shown in Table 5, we can observe that $D = 8$ can get the relatively best results. Compared with the baseline module without the part guidance, there is a consistent improvement with the detected part regions. It shows that these part regions are able to narrow down the possible searching area, which is helpful for focusing on the valid part components. We also observe that our PGAN can gradually improve the Re-ID performance with the number of part regions increasing. However, the performance has a limitated increase with the part number. The reasons are twofold: 1) a lot of detected part regions are covered with each other, which provide no further part information; 2) more wrongly detected parts with invalid information are extracted that might result in the distraction for model learning. We believe that if we use a better detector, the performance can be further improved.

Part Number D	Dimension	mAP	Top-1	Top-5
0 (Baseline)	512	75.2	94.9	98.1
4	512	76.4	74.4	97.9
6	512	78.7	96.2	98.0
8	512	**79.3**	**96.5**	**98.3**
10	512	79.1	95.9	98.2

Table 5: Performance comparison on different part number D of PGAN on VeRi-776.

A.3 Qualitative Analysis of the Performance

In this section, we visualize more retrieval results of the grid attention and our part-guided attention (PGAN), repectively. As illustrated in Figure 5, we illustrate four different query vehicle images and their corresponding Top-5 most similar images as well as the heatmaps of the part-guided feature $\mathbf{F}_p$ from PAM in the gallery set. From Figure 5, we observe that our PGAN is able to obtain more reliable retrieval results compared with the grid attention method. In detail, the main advantages of our PGAN can be summarized as follows:

1. **Insensitive to Various Situations.** The PGAN can extract more robust feature representation so as to significantly improve the Re-ID performance. As shown in the ID1 and ID4, given a rear vehicle image, we can not only find the easy vehicles from the rear and side views, but also get the front-view vehicle images that are difficult to recognize even by humans. In contrast, the grid attention can only focus on the images from the nearly same views. Moreover, our PGAN is also able to deal with various situations, such as illumination and occlusion. It means that our method is more robust to learn discriminative features that is not sensitive to multiple variants from the environment.

2. **The Effectiveness of the Part Extraction Module.** The detected part regions play an important role in the feature representation. As illustrated in the ID3, it is clear that the wrongly retrieved images from the grid attention method have different car lights with the query image. However, a lot of regions representing the body and the bottom of the car are concentrated, which are not the obvious differences between two vehicles. With the guidance of the detected part regions, our PGAN can only concentrate on these candidate regions, *e.g.*, car lights in the ID3. It helps the model focus on the useful regions as well as alleviating the bad effect from the other regions. That is to say, the part extraction module is beneficial for the network learning by narrowing down the searching areas.

3. **The Effectiveness of the Part Attention Module.** Our PGAN is useful for selecting the most prominent part regions and lighten the influence of invalid and useless regions. As described in the main paper, we propose a Part Attention Module (PAM) that is responsible for learning a soft attention weight for each part. Therefore, the important part regions are underlined by a high-attention value, while the impact of other insignificant parts is relatively suppressed. From the heatmaps, we can clearly observe that our PAM could focus on the most significant part regions, such as the car lights in ID3, back mirrors in ID1. As shown in ID4, although there are few valid part regions that are extracted, our PAM can still find the key information to recognize the vehicles,

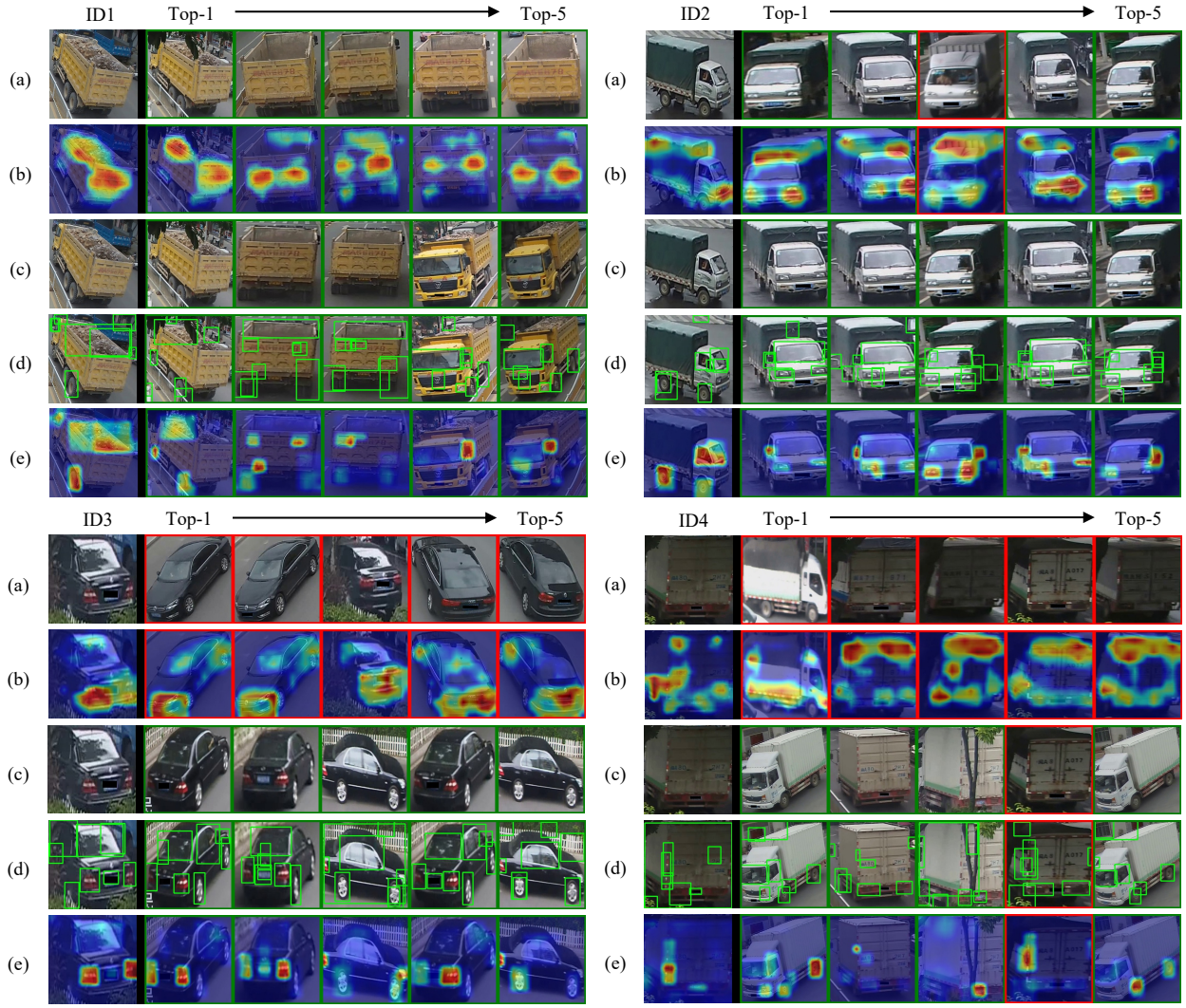

Figure 5: Visualization of the retrieval results of the traditional grid attention and our PGAN. We illustrate 4 vehicle images with the most Top-5 similar vehicles in the gallery set. The correct and false matched vehicle images are enclosed in green and red rectangles respectively. For a query image, we draw: the results of the grid attention in (a) the retrieved vehicle images and (b) the corresponding heatmaps of the part-guided feature $\mathbf{F}_p$ from PAM; and the results of our PGAN in (c) the retrieved vehicle images and (d) the detected candidate part regions and (e) the corresponding heatmaps of $\mathbf{F}_p$. (Best viewed in color)

such as the wheel and the car lights. On the contrary, the grid attention is largely influenced by some invalid regions with extremely similar appearance in different vehicles, such as the bottom of the vehicle body.

References

[1] C. Arth, C. Leistner, and H. Bischof, "Object reacquisition and tracking in large-scale smart camera networks," in *Proc. ACM/IEEE Int. Conf. Distributed Smart Cameras*, pp. 156–163, 2007.

[2] R. S. Feris, B. Siddiquie, J. Petterson, Y. Zhai, A. Datta, L. M. Brown, and S. Pankanti, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 28–42, 2012.

[3] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3973–3981, 2015.

[4] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," pp. 1–6, 2016.

[5] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 379–387, 2017.

[6] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. AAAI Conf. Artificial Intell.*, 2018.

[7] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 8797–8806, 2019.

[8] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3235–3243, 2019.

[9] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comp. Vis.*, pp. 869–884, 2016.

[10] 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 *Proc. Asian Conf. Comp. Vis.*, pp. 575–591, 2018.

[11] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2017.

[12] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 1900–1909, 2017.

[13] Y. Zhou and L. Shao, "Cross-view gan based vehicle generation for re-identification.," in *Proc. British Machine Vis. Conf.*, vol. 1, pp. 1–12, 2017.

[14] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 6489–6498, 2018.

[15] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Transactions on Image Processing*, 2019.

[16] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3997–4005, 2019.

[17] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Vehicle re-identification using quadruple directional deep learning features," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[18] Y. Zhao, C. Shen, H. Wang, and S. Chen, "Structural analysis of attributes for vehicle re-identification and retrieval," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[19] H. Chen, B. Lagadec, and F. Bremond, "Partition and reunion: A two-branch neural network for vehicle re-identification," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 184–192, 2019.

[20] Y. Chen, L. Jing, E. Vahdani, L. Zhang, M. He, and Y. Tian, "Multi-camera vehicle tracking and re-identification on ai city challenge 2019," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 324–332, 2019.

[21] X. Liu, S. Zhang, Q. Huang, and W. Gao, "Ram: a region-aware deep model for vehicle re-identification," pp. 1–6, 2018.

[22] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 6077–6086, 2018.

[23] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 2167–2175, 2016.

[24] A. Kanacı, X. Zhu, and S. Gong, "Vehicle re-identification in context," in *Proc. German Conf. Pattern Recognition*, pp. 377–390, 2018.

[25] Y. Tang, D. Wu, Z. Jin, W. Zou, and X. Li, "Multi-modal metric learning for vehicle re-identification in traffic surveillance environment," in *Proc. IEEE Int. Conf. Image Process.*, pp. 2254–2258, 2017.

[26] X. Liu, W. Liu, H. Ma, and S. Li, "A progressive vehicle search system for video surveillance networks," in *Proc. IEEE Int. Conf. Multimedia Big Data*, pp. 1–7, 2018.

[27] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 562–570, 2017.

[28] A. Sanakoyeu, V. Tschernezki, U. Buchler, and B. Ommer, "Divide and conquer the embedding space for metric learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 471–480, 2019.

[29] R. Kumar, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: an efficient baseline using triplet embedding," *arXiv preprint arXiv:1901.01015*, 2019.

[30] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 814–823, 2017.

[31] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual path modelwith adaptive attention for vehicle re-identification," *arXiv preprint arXiv:1905.03397*, 2019.

[32] A. Kanaci, M. Li, S. Gong, and G. Rajamanoharan, "Multi-task mutual learning for vehicle re-identification," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 62–70, 2019.

[33] P. Khorramshahi, N. Peri, A. Kumar, A. Shah, and R. Chellappa, "Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 239–246, 2019.

[34] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[35] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. Workshops of IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019.

[36] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," *arXiv preprint arXiv:1904.01400*, 2019.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 7132–7141, 2018.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 770–778, 2016.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 248–255, 2009.

[40] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.