Vehicle Re-Identification Using Machine Learning

Lisha Tang

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Master of Engineering

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

29-07-2021
.....
Date

.....
Lisha Tang

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

29-07-2021

.....

Date

.....

Prof. Lap-Pui Chau

Authorship Attribution Statement

This thesis contains material from one manuscript submitted in the following peer-reviewed journal in which I am listed as an author.

Most parts of this thesis are submitted and under review in IEEE Transactions on Circuits and Systems for Video Technology as Lisha Tang, Yi Wang, and Lap-Pui Chau, "Looking Twice for Partial Clues: Self-supervised Part-Mentored Attention Network for Vehicle Re-Identification", 2021.

The contributions of the co-authors are as follows:

- I came up with the original idea of this work, implemented the experiments and wrote the first manuscript draft.
- Dr. Yi Wang provided many valuable recommendations for this work and revised the whole manuscript draft.
- A/Prof Chau provided the initial project direction, did the quality control of the draft and revised the entire manuscript draft.

29-07-2021

.....

Date

.....

Lisha Tang

Acknowledgements

I wish to express my greatest gratitude to my supervisor, Prof. Lap-Pui Chau, for his patient guidance. Thank him for providing the research direction, sharing his valuable experiences, and giving me encouragement as well as recommendations during my graduate study.

Besides, I would like to appreciate all the group members for their help. I am grateful to my colleague, Dr. Yi Wang, for giving me precious suggestions for my work. He is a competent mentor with rich research experience.

Finally, I would like to express my sincere appreciation to my family and friends for their continuous support. Thank Kening Sun for encouraging and comforting me.

Abstract

Vehicle Re-ID aims to retrieve images of the same vehicle across non-overlapping cameras. The key challenges lie in the subtle inter-class discrepancy caused by near-duplicated identities and the significant intra-class distance due to diverse factors, including illumination, viewpoints, and background interference. This thesis starts with reviewing the development history of vehicle Re-ID and proposes a Part-Mentored Attention Network (PMANet) consisting of a Part Attention Network (PANet) for weakly-supervised vehicle part localization and a Part-Mentored Network (PMNet) for mentoring the global and local feature aggregation. Firstly, PANet predicts a foreground mask and pinpoint K prominent vehicle parts without additional part-level supervision. Secondly, PMNet applies multi-scale soft attention on localized regions and compensates inaccurate part masks with part-guided learning. PANet and PMNet construct a two-stage attention structure to perform a coarse-to-fine search among identities. Finally, we address this Re-ID issue as a multi-task problem and employ Homoscedastic Uncertainty Learning to automatically balance the loss weightings. Experimental results show that our approach outperforms recent state-of-the-art methods by averagely 2.63% in CMC@1 on VehicleID and 2.2% in mAP on VeRi776.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter starts with task definition and motivation, explains our objectives and major contributions, and then briefly introduces the overall organization of this thesis.

1.1 Motivation

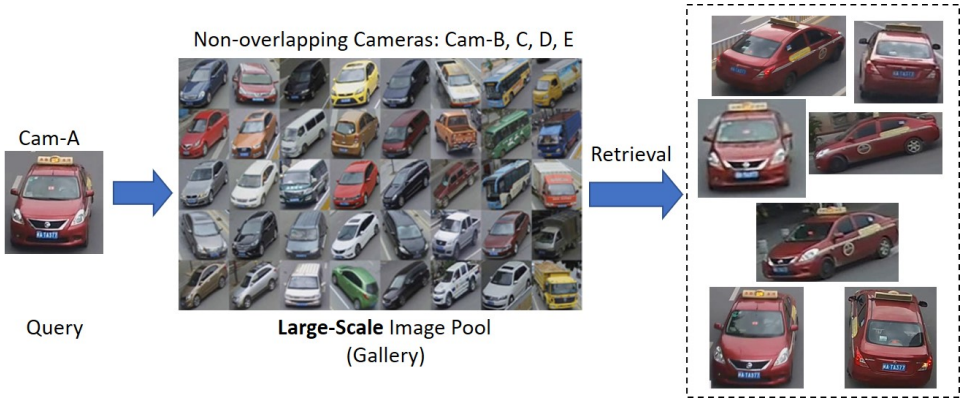

FIGURE 1.1: Definition of vehicle re-identification.

Given a query vehicle image, vehicle re-identification (Re-ID) aims at identifying all images with the same identity in a gallery set across multiple non-overlapping cameras. It holds great potential in public security as well as intelligent transportation, thereby drawing increasing attention from both academia and industry. Despite this, various challenges still hinder the performance of vehicle Re-ID (see Fig.1.2a): (1) the large intra-class variance: a vehicle may exhibit dramatically different visual appearances due to the large number of uncontrolled variation sources, such as illumination, viewpoint

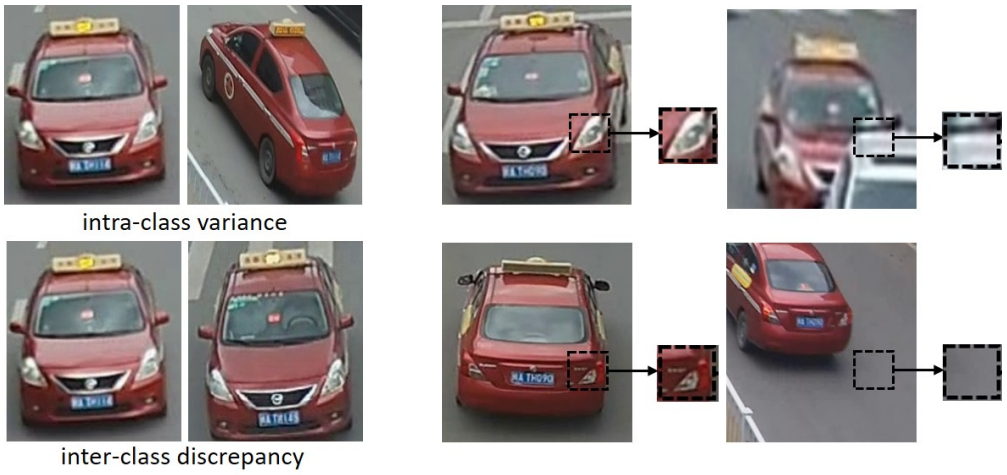

FIGURE 1.2: (A) The two images in the first row show the large intra-class variance of the same vehicle caused by viewpoint changes, while the images in the second row illustrate the minor inter-class discrepancy caused by near-duplicated vehicles. (B) The four zoomed-in patches show that the same spatial position across images may not correspond to the same vehicle part due to background interference, viewpoint variation and translation.

changes and poor image quality; (2) the subtle inter-class discrepancy: different vehicles with the same model and color may look quite similar especially when the images are captured from one single unified view.

The key to robust ReID lies in multi-granularity features. Recent works elaborately aggregated local features with global ones to enhance feature representations. Three main weaknesses are observed among the state-of-the-art approaches. Firstly, one intuitive strategy is to extract local features from several vertically separated vehicle regions [37, 47]. However, as shown in Fig.1.2b, such naive uniform division fails to handle spatial component misalignment because the same spatial position across images may not correspond to the same vehicle part owing to viewpoint changes, object translation or background interference. In extreme cases, uniform division might miss some crucial information, thus leading to performance degradation. Other methods mitigate such misalignment by parsing images into semantic-aware or view-aware masks with outside tools [43, 14, 47, 13, 3, 79, 77, 57]. For example, He *et al.* [14] pre-defined windows, lights and brands for each vehicle and used them to train a YOLO [50] detector. PVEN [43] parses each image into four views based on the 20 key-points labelled. Nevertheless, these methods require substantial pixel-level annotations, which is time-intensive and labor-intensive. Thirdly, existing global-local feature fusion methods [64, 14, 3, 43,

13] closely rely on their generated part masks. However, these masks are not always accurate (see Fig.1.3), thus introducing noises to the model.

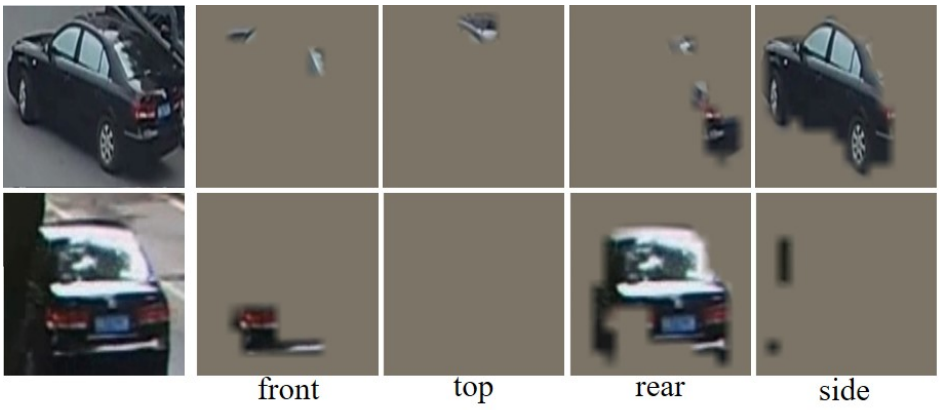

FIGURE 1.3: The figures show the four view-aware masks of two sample images generated by PVEN [43].

1.2 Objectives and Major Contributions

The objectives of this thesis are two-fold. To begin with, we aim to conduct thorough research on vehicle Re-ID to gain basic knowledge and a high-level perspective of current obstacles in this issue. Then to handle the aforementioned weaknesses in Chapter 1.1, this thesis proposes a Part-Mentored Attention Network (PMANet) consisting of a weakly-supervised Part Attention Network (PANet) and a Part-Mentored Network (PMNet), as shown in Fig.3.1. To address spatial misalignment without additional annotations, PANet predicts a refined foreground mask and robustly locates K prominent vehicle parts. Afterwards, PMNet compensates inaccurate part masks with part-guided learning. Specifically, for each part-level feature learning branch, PMNet applies a noisy teacher to guide the learning of a student so that they can concentrate on the same vehicle part. During testing, the noisy teacher can be disgarded, which means PMNet is able to complete global-part feature aggregation without prior mask prediction. Briefly, our contributions can be summarized as follows:

1) To our knowledge, this is the first attempt to locate different vehicle parts using attention clustering without extra partial supervision in vehicle Re-ID. Our PANet is simple in structure and easy to optimize compared to other attention-based methods [38, 31].

2) We propose a teacher-student guided learning structure in PMNet to learn a more robust feature representation in case of inaccurate part masks which lead to the failure of state-of-the-art methods[14, 3, 43, 77].

3) An end-to-end multi-task learning scheme is introduced to train PMNet, thus improving the model generalization and performance. Besides, Homoscedastic Uncertainty Learning [24] is adopted to expedite the convergence process and learn loss weighting automatically.

4) Extensive experiments and ablation studies are conducted to demonstrate the superiority of our method compared with recent approaches. Results on occluded test sets also prove that our method effectively resists the obstruction in the occluded vehicle images.

1.3 Outline of the Thesis

Chapter 2 reviews related literature and briefly analyzes their advantages as well as disadvantages.

Chapter 3 describes the detailed structure of our proposed method, Part-Mentored Attention Network (PMANet).

Chapter 4 presents implementation details, dataset information and experiments which we conduct to investigate the effectiveness of our method.

Finally, in Chapter 5, we make a conclusion, followed by some recommendations for further research.

Chapter 2

Literature Review

2.1 Vehicle Re-ID

2.1.1 Hand-crafted feature based methods

Derived from person re-identification (Re-ID), vehicle Re-ID has attracted increasing attention, but the performance is still unsatisfactory. Early vision-based vehicle re-ID works leveraged hand-crafted methods for feature extraction [65, 81, 74]. Among them, [65] proposes a weighted approximate string matching to improve the licence plate matching performance. [81] estimates invisible attributes with visible ones via query expansion and re-rank results by employing special attributes. [74] employs color histogram and oriented gradient histogram followed by linear regressors. However, the performances of such methods are limited and licence plates are not always available for vehicle Re-ID. Based on these limitations, [39] combines hand-crafted features, like Color Name and SIFT, with deep features extracted from CNN models. With recent breakthroughs of deep learning, experiments demonstrate that deep features are more discriminative and effective than hand-crafted ones. From then on, many deep learning-based models have dominated vehicle Re-ID.

2.1.2 Deep feature based methods

Vehicle Re-ID aims to find all the images of the same vehicle that appears at different time and locations in the large-scale camera network, in which feature representation

and distance metric learning are two fundamental elements. Therefore, recent deep features based works are generally categorized into two groups. One group concentrates on designing metrics for similarity measurement between images captured in different views, while the other makes efforts in developing comprehensive and effective feature representations.

2.1.2.1 Metric learning

As regards the first category, the minor inter-instance discrepancy and large intra-instance variance relate to two basic obstacles in Re-ID tasks. To cope with these problems, many metric learning-based works aim at learning a feature embedding space that maximizes inter-instance distances and simultaneously minimizes the intra-instance distances. Based on the lifted dense pairwise distance matrix in the training batch, [45] reduces the distances between similar sample pairs while maximising the distances between dissimilar ones. DRDL [35] proposes a Coupled Clusters Loss (CCL) for feature representation enhancement. [67] designs a new framework that improves kNN classification by incorporating a Mahalanobis distance metric.

Particularly, [19] introduces a Triplet loss which learns a feature embedding such that samples with the same identity are closer to each other than those with different identities. Such triplet constraints have been widely utilized and adapted in face recognition [53] as well as person Re-ID [20, 4, 71, 70] tasks. Based on the triplet loss, [5] also develops a quadruplet network to improve the generalization capability of learnt feature representations. Lin *et al.* [85] incorporated bipartite-graph labels (BGL), which models the rich relationships among the ultra-fine grained classes, into an overall convolutional neural network. Zhang *et al.* [48] proposes a multi-stage metric learning network to explicitly address the problems caused by too many triplet constraints, such as storage limitation and computational challenge. [76] seamlessly embeded label structures such as hierarchy (e.g., make, vehicle model) or attributes (e.g., ingredients of food) into feature learning as prior knowledge, and jointly optimized the classification loss (i.e., softmax) as well as the similarity loss (i.e., triplet). Yang *et al.* [72] proposed a self-trained subspace learning framework which uses both labeled and unlabeled samples to construct a discriminant metric. It is observed in [7] that by simply using the common triplet loss, instances belonging to the same identity might make up a large cluster with a relatively large average intra-class distance in the learnt feature space. Therefore, to further constrain the intra-instance distance, [7] improves the triplet loss with a new term.

Besides, [68] proposes a new supervision signal, named the center loss. This loss computes the center of feature representations for each identity and penalizes the distances between each feature and its correspondent identity center. Moreover, some works fuse semantic knowledge into metric learning. [33] jointly exploits visual features and the user-provided tags to learn a distance metric, which can preserve the semantic structure and the visual structure simultaneously. Cui *et al.* [9] formulated a novel Knowledge Embedded Representation Learning framework that incorporates high-level knowledge graph as additional guidance for feature learning. With the guidance, this framework learns feature maps with a meaningful configuration that the highlighted regions are finely related to the attributes in the knowledge graph.

For vehicle Re-ID task, Apart from constructing a large-scale vehicle database named PKU-Vehicle, Bai et al. [2] designed a group-sensitive-triplet embedding that injects the inter-class discrepancy and the intra-class variance into a triplet embedding. Kumar et al. [29] summarized and compared different batch sampling ways for triplet loss. Nevertheless, all these metric learning-based methods merely consider the distance between holistic features but ignore part-level features, which may provide crucial clues especially between near-duplicated identities.

2.1.2.2 Feature representation

Attribute-based feature learning Early feature learning based methods directly incorporated various meta information such as vehicle attributes (e.g., model, color) and spatial-temporal information to enhance global feature representation. [84] fuses camera views, vehicle types and color into feature embedding. DJDL [32] proposes a unified framework which efficiently optimizes multiple tasks, including attribute recognition, identification, triplet and verification tasks. The entire framework is jointly optimized by a batch composition design. [55] reduces searching space by utilizing visual-spatial-temporal paths. However, these global representation-based approaches suffer from the instability caused by dramatic viewpoint changes and challenges brought by the large intra-class variance. Besides, attributes and spatial-temporal cues are not always available, limiting the application of those algorithms. In contrast, our work does not consider the usage of additional information or annotations.

Regional feature learning To learn more discriminative details that holistic features fail to provide, recent works take local feature learning into account. Current regional feature learning methods can be grouped into three categories. The first category [47, 37] utilizes uniform division, which vertically splits the image or feature map into several regions of equal area. However, as is mentioned in the Introduction, such naive division suffers from spatial component misalignment due to viewpoint changes, occlusion and object movement. Others [14, 47, 13, 3, 43, 79, 77, 57] mitigate such issue by pre-defining vehicle components with outside tools. [79] collects a dataset with 21 classes of attribute labels and trains an adapted Single-Shot Detector [36]. PGAN [77] detects vehicle components based on [79] and proposes an attention module which adaptively estimates the importance of detected components. This attention module assigns high weights to the most discriminative regions and applies relatively low weights to irrelevant parts. PRReID [14] detects windows, headlights, and brands for each vehicle by training a YOLO [50] detector. The third type learns viewpoint-specific features. For example, Wang et al. [64] labeled 20 vehicle key-points and predict attention maps by dividing the key-points into four categories which represent front, left, right or rear view of a vehicle, respectively. MVAN [59] proposes a multi-view branch network, in which each branch concentrates on a limited range of viewpoint variations. However, all these methods mentioned above require additional data annotations. Besides, TAMR [13] leverages two STNs [22] to locate windscreens and car heads, and constrains its model with a Multi-Grain Ranking loss. Apart from requiring central coordinates of these two vehicle components, TAMR also ignores other vehicle parts, which might also contain important local information (e.g., front bumper, tires). It is also notable that headlights and windscreen cannot be captured under some views, such as side view. Therefore, self-supervised and weakly-supervised vehicle part localization is still challenging.

Some other localization methods, e.g., SCDA [66], PL-Net [73] and PAN [83], are proposed based on the principle that different channel groups tend to exhibit strong activation on regions with different semantic meanings. PAN [83] re-locates the pedestrian with STN [22] and merges it with the feature of the original image to form a new pedestrian descriptor. However, this method merely concentrates on re-localization of the full human body but fails to take advantage of the fine-grained details which can be provided by different body parts. PL-Net [73], which clusters the coordinates with the maximum response of each channel for component localization, is proved quite unstable because without further channel grouping, feature maps of the encoder are sensitive to inference (e.g., noises and background clutter), and pixels with the maximum response might be

noise or lie in the same body part of a person. Differently, our method can perform vehicle part localization robustly.

Other vehicle Re-ID methods Some researchers also design models based on Long Short-Term Memory (LSTM) and Generative Adversarial Networks (GAN) in vehicle Re-ID. For instance, Zhou *et al.* [86] exploited the great advantages of the LSTM to model transformations across continuous view variations of a vehicle. As for GAN, Lou *et al.* [40] proposed a GAN-based model for hard sample generation. Zhou *et al.* [87] designed a viewpoint-aware GAN-based network which generates multi-view features by merely using a single-view vehicle image. Nevertheless, due to the insufficient adversarial samples and the limited generation ability of existing GAN, there still exists a large gap between the generated samples and real ones. In addition, transfer learning is recently introduced to person Re-ID [78, 51, 10]. A DSAG-Stream designed in [78] guides an MF-Stream to learn densely semantically aligned features. [10] handles body part misalignment through multi-task learning in a student-teacher manner. However, relevant exploration in vehicle Re-ID is quite insufficient. Inspired by [10], our PMNet aligns global and local features in part concept transfer manner to enlarge inter-instance discrepancy especially between similar cases.

2.1.3 Occluded Re-ID

Different from traditional Re-ID task which performs pedestrian or vehicle retrieval in the full-body domain, occluded Re-ID is a challenging real-world issue. Re-ID suffers from a serious occlusion problem especially when applied to crowded public places. As is shown in Fig.2.1, there are two main challenges for occluded Re-ID. Firstly, with various obstruction regions, images are less discriminative, making it even harder to distinguish similar identities. Moreover, traditional methods can easily regard the occlusion as the texture of the person or vehicle, thus corrupting the representation. Secondly, since occlusion may occur in any area, it is difficult to locate different parts without prior alignment information. So far, occluded person Re-ID has attracted much attention [82, 16, 18, 17, 75, 56, 88, 42]. For instance, [42] employs an affine transform model that transforms the global image and aligns it with the occluded ones. [75] boosts its model performance by locating the occlusion region and recovering the occluded person. Zhou *et al.* [88] introduced an attention framework, which automatically generates

artificial occlusions for full-body person images and optimizes multi-task losses. To handle occluded Re-ID without additional supervision, Zheng *et al.* [82] proposed a local patch-level matching model and a global part-based matching model which supplies complementary spatial alignment clues. [56] designs a self-supervised model to locate the visible regions.

Nevertheless, there is still little research related to occluded vehicle Re-ID. [34] increases the model generalization by synthesizing occlusion samples and resorts to an attention mechanism for fine-grained feature learning. ASAN [23] proposes a CAM-based segmentation module and a shift feature adaptation module to extract features within the visible part of the image. It designs an occluded test set based on VeRi776 [39], and conduct comparison experiments on it. However, this method uses vehicle model and color information which are not always available.

FIGURE 2.1: The two sub-figures show samples for occluded person Re-ID and vehicle Re-ID, respectively. In each sub-figure, images on the first row denote the whole-body images, while those on the second row indicate the corresponding occluded samples.

2.2 Attention Mechanism

Attention mechanism in deep learning, which mimics the human visual attention mechanism [12], has been adopted in many vision tasks to mine more discriminative information. To capture channel-wise dependencies, SE-Net [21] compresses global spatial information into a channel descriptor. CBAM [69] proposes a lightweight module which learns attention maps in both spatial and channel dimension. All these methods concentrate on a global scale. Recently, researchers have realized that the scale variation of objects plays an essential role in many fine-grained tasks and introduced multi-scale attention to handle such issue in CNNs. Multi-scale attention mechanisms are mostly

achieved by either mixing feature contexts with various scales inside an attention module, or inputting multi-scale images or features into the module. The first type leverages multiple convolution layers [49] or a pyramid [63] for feature context aggregation inside the attention module. The second type feeds features at multiple scales or their concatenated representation into the attention module to produce attention maps. Still, the scale of feature context aggregation inside the module stays single [8, 30].

Meanwhile, high-order attention methods are proposed in video classification, visual question answering (VQA), and fine-grained visual categorization (FGVC). Specifically, [27] designs a bilinear attention module to handle the relationship between image regions and the words in question, and [80] introduces a third-order operation to get inter-relationship among channels.

Recent vehicle Re-ID works also introduce attention modules [60, 77, 13, 59]. Among them, Teng *et al.* [77] developed a channel-wise Part Attention Module to indicate the importance of each part feature. Nevertheless, research in multi-scale attention is still limited and all these methods employ attention either for detailed part clues or for weighting of their extracted features. In this project, our attention method differs from them in the following aspects: 1) We introduce a hard part-level self-attention network, PANet, to suppress noises of the feature maps and locate informative vehicle parts. 2) The Multi-Scale Attention module applies pixel-level soft attention on the localized vehicle parts in both spatial and channel dimensions. 3) Such two-stage attention structure gives a second look at the refined region and helps mine features of multiple granularities.

2.3 Multi-task Learning

Multi-task learning aims at simultaneously optimizing multiple relevant tasks under a shared model. It can be thought of as an inductive knowledge transfer method that enhances generalization by exchanging domain information between complementary tasks. A variety of deep learning applications benefit from multi-task learning [1, 46, 44, 25, 58, 11, 61, 28]. For instance, Cross-Stitch [44] combines the activations from multiple networks. PoseNet [25] learns camera position and orientation. To enable real-time application, MultiNet [58] combines detection, classification and semantic segmentation in

a unified architecture. [11] addresses three sub-tasks, containing surface normal estimation, semantic labeling and depth prediction, together by using an end-to-end multi-scale convolutional network. Multi-task learning has also been introduced into regression and geometry tasks. Low-, mid-, and high-level vision tasks are jointly optimized in a unified architecture in [28]. Such multi-task solutions help improve data efficiency, reduce overfitting through shared representations, and increases learning speed by utilizing auxiliary information.

Moreover, it is observed that the performance of such approaches relies on the weighting ratio between the losses. Prior works simply conduct a weighted linear summation on the losses for each task, where the loss weights are uniform or manually tuned [54, 28, 11]. However, manual tuning for an optimal weighting is costly, especially when the number of tasks is above 2. In our proposed method in Chapter 3, we model this vehicle Re-ID issue as four sub-tasks and adopt HUL [24] to learn the optimal loss weighting.

It is observed in [24] that the optimal task weighting relies on the measurement scale (e.g. millimeters, centimeters or meters) and eventually the magnitude of each task's noise. HUL [24] treats the multi-task network as a probabilistic model and derives a weighted multi-task loss function by maximizing a Gaussian likelihood objective which accounts for the homoscedastic uncertainty. Note that, the weight for task i is inversely proportional to the introduced noise parameter σ_i. As a result, the higher the task's homoscedastic uncertainty is, the smaller the influence of task i on the network weight update gets. This is favorable when coping with noisy labels because the task-specific weights for such tasks will be automatically decreased.

Chapter 3

Methodology

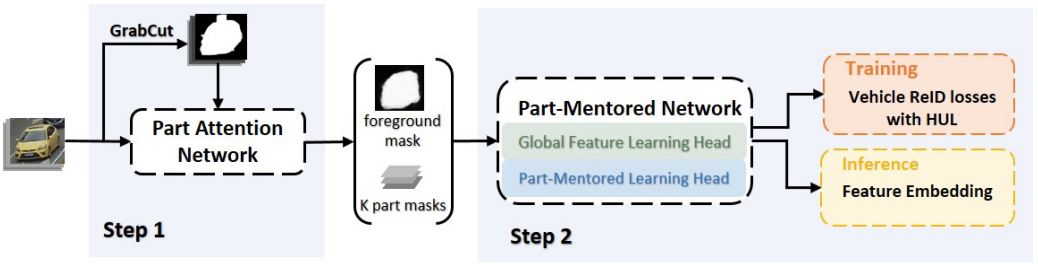

FIGURE 3.1: Pipeline of our method. Our proposed method consists of two steps, Part Attention network (PANet) and Part-Mentored Network (PMNet). Utilizing coarse masks by GrabCut [52] as pseudo labels, PANet refines the foreground mask and predicts K part masks without part-level supervision. They are fed into PMNet for global and local feature learning.

In this chapter, we will present the details of our proposed method, Part-Mentored Attention Network (PMANet). When differentiating objects, humans first utilize obvious general information for easy samples, then observe salient components closely for more discriminative clues to distinguish near-identical cases. Our method mimics such human strategy and performs a coarse-to-fine search on vehicles. Typically, vehicle part location information is not supplied by datasets. Therefore, our model learning is weakly supervised in the context of optimizing Re-ID performance. As illustrated in Fig.3.1, our proposed method consists of two steps: a weakly-supervised Part-Attention Network (PANet) in Chapter 3.1 and a Part-Mentored Network (PMNet) in Chapter 3.2. In Chapter 3.3, we model the entire Re-ID issue as four sub-tasks and incorporate Homoscedastic Uncertainty Learning to automatically balance the joint learning of the ID losses.

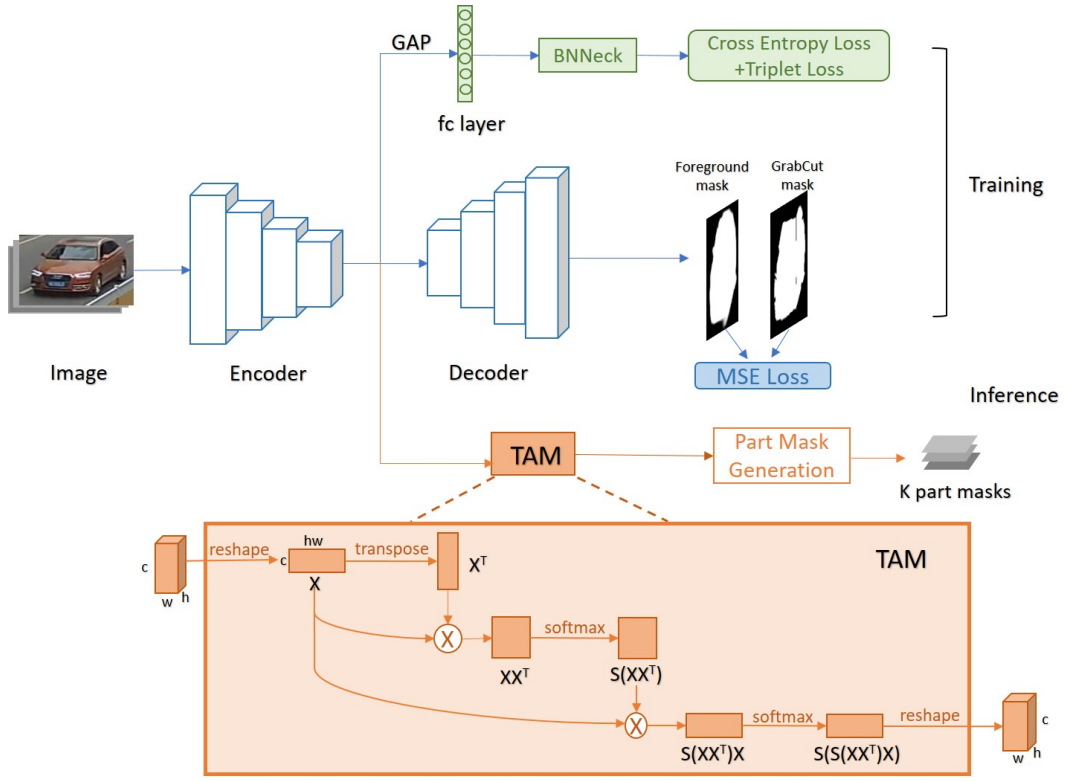

FIGURE 3.2: Architecture of the Part Attention Network. As the blue arrow lines show, during training, the PANet combines a classic Re-ID baseline model [41] and an encoder-decoder style segmentation model to learn discriminative features for identity classification and reconstruct a foreground mask at the same time. The orange arrow lines illustrate the pipeline to predict K part masks in the inference stage and details of the Trilinear Attention Module (TAM) are shown in the orange rectangle. $\mathcal{S}(.)$ in TAM indicates softmax normalization and $\otimes$ denotes matirx multiplication.

3.1 Part Attention Network for Vehicle Part Localization

Part Attention Network (PANet) aims at automatically locating K part masks which focus on different vehicle parts. Inspired by [3], we first generate coarse foreground masks of vehicles using the handcrafted segmentation algorithm, GrabCut [52], without any annotation. However, these masks are unstable and often contain errors, thereby are used as pseudo labels for PANet.

As shown in Fig.3.2, our PANet combines a classic classification and segmentation model in a unified framework to extract discriminative features for identity classification and foreground mask refinement. During training, our PANet employs ResNet-50

[15] as the feature encoder. In the classification branch, the feature map from the encoder is fed into a fully connected layer, a BNNeck [41] to compute ID loss and triplet loss [19]. In the segmentation branch, four blocks are introduced as the decoder to calculate the Maximum Squared Error (MSE) loss between its reconstructed masks and coarse pseudo masks. In specific, the first three blocks are composed of one transposed convolutional layer, a batchnorm layer and a ReLU layer, while the last one consists of a transposed convolutional layer and Sigmoid. The segmentation branch considers vehicle structure to help inform class boundary decisions, while the classification branch introduces instance-specific supervision and helps weigh the importance of different vehicle parts. Such a combination of the classic classification as well as segmentation framework leads to a compromise between the classification task and the segmentation task under the same backbone. It simultaneously generates a refined foreground mask and an intermediate feature map that weighs the importance of different vehicle parts. During inference, the output feature maps from the encoder are further fed into a Trilinear Attention Module (TAM) and Part Mask Generation module to predict K part attention masks.

Recent works in Person Re-ID have shown that different channel groups of the feature map for a person exhibit strong responses at their corresponding body parts [73]. Similar correspondences between channels and vehicle parts are also observed in vehicle Re-ID. As depicted in Fig.3.3, although information conveyed by each single channel is weak and sensitive, many channels usually tend to fire on the same region which may indicate different semantically meaningful parts (e.g., car roof, bonnets, windscreens and headlights). This property can assist in locating different vehicle parts in the spatial dimension, thereby learning detailed local information within each vehicle part. However, some channels might be sensitive to noise and activate at the background or other wrong spots, creating a negative impact on the overall performance (e.g., PL-Net [73]). Therefore, it is important to design a network which helps accurately group different channels and suppress noise. With this goal, we propose a Trilinear Attention Module to convert the feature map to a more robust and consistent attention map, and then the map is fed into a Part Mask Generation module to locate partial masks.

As Fig.3.2 depicts, during inference, given an input image, we obtain the feature map from the encoder with a dimension of $C \times H \times W$, where C, H and W respectively indicate the channel number, height, and width. After reshaping the feature map into

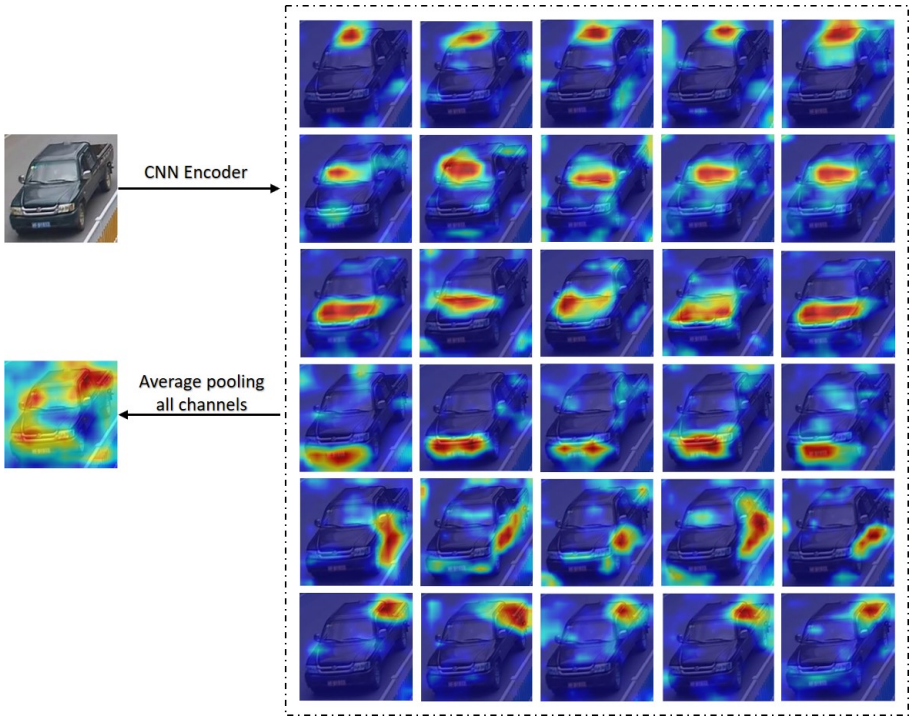

FIGURE 3.3: Examples of attention maps of the output feature channels generated by the CNN encoder.The two figures on the left are the input vehicle image and the attention map after average pooling all the channels. The figures on the right are respectively attention visualizations of channels which indicate different vehicle parts, e.g., vehicle roof, windscreens and head lights. The images on each row exhibit channels that fire on the same region.

a matrix, $X \in \mathbb{R}^{C \times HW}$, we compute the bilinear feature, $G(X) = softmax(X \cdot X^T)$, which indicates the relationship between the channels.

Since channels within a channel group concentrate on the same part, $G(X)$ can group the channels of the feature map with a bilinear product and the $softmax$ operation employed here computes the similarity between any two channels. The larger $G_{i,j}(X)$ is, the more similar channel i and channel j are, i.e., the more likely channel i and channel j concentrate on the same vehicle part. Then, we get a more consistent feature map by multiplying $G(X)$ with the reshaped feature map, X. In this way, for each channel, important information about the vehicle region is amplified by other channels within the same group, while noisy information is reduced. Another $softmax$ is conducted on the output feature map to generate a normalized attention map. This trilinear function can be formulated as,

$$F(X) = \mathcal{S}\left(\mathcal{S}\left(X \cdot X^T\right) \cdot X\right), \tag{3.1}$$

where $\mathcal{S}(.)$ indicates softmax normalization over the second dimension of the matrix. Finally, the attention map is reshaped to $C \times H \times W$. In brief, this self-trilinear product suppresses influence of background clutter as well as random noise, conducts further channel grouping, and generates a more stable attention map.

Afterwards, the generated normalized attention maps are fed into the Part Mask Generation module to predict partial masks. In specific, we first obtain the bounding box of the largest connected domain of each channel, calculate the respective central coordinate and conduct k-means clustering on all of them according to their positions in each feature map. K cluster centers are obtained and thus, K rectangular partial masks are gained. Here, for each partial mask, we set the height $h = \frac{H}{K}$; $w = \frac{W}{2}$.

Remarks Our hard attention modeling is conceptually similar to some weakly-supervised localization approaches [66, 73, 83] but the design differs in the following aspects: $(\mathbf{1})$ PANet performs simple attention clustering on a more robust and consistent vehicle feature map, which saves model parameters and is easy to optimize. $(\mathbf{2})$ These methods merely consider region-level hard attention whilst our approach exploits complementary benefits by constructing soft attention modeling within selected regions at the pixel level.

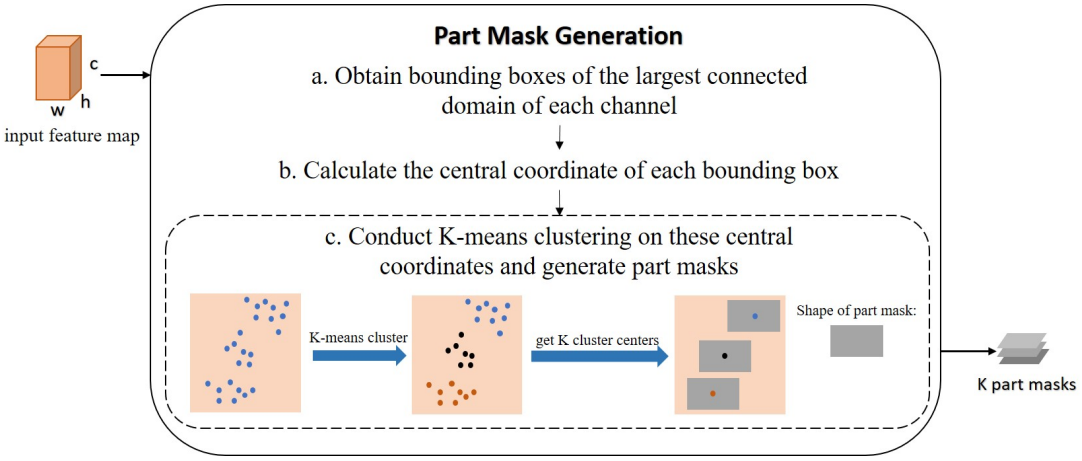

FIGURE 3.4: Details of Part Mask Generation module.

3.2 Part-Mentored Network for Vehicle Re-ID

In this section, we first describe the motivation and problem formulation of vehicle Re-ID. Then, the details of our Part-Mentored Network (PMNet), including global feature learning head and the Part-Mentored Learning Head (PMLH), are presented. The overall architecture is depicted in Fig.3.5.

3.2.1 Problem Formulation for Re-ID

Given a query image I_q, vehicle Re-ID seeks to retrieve the images with the same identity from a gallery set G. The gallery set is denoted as $G = \{I_i\}$, $i \in [1, m]$, where m is the total number of images in this gallery set. This issue is usually solved by learning a distinguished feature representation, f, for each image from a training set. Thereby, the retrieval process can be simplified as matching the feature representation of the query, f_q, with those of the gallery images, f_G.

Suppose the training set contains N labeled images from T vehicles, we indicate the training set as $D = \{I_i, y_i\}$, $i \in [1, N]$, $y_i \in [1, T]$, where I_i denotes the i-th image and y_i is the correspondent vehicle identity. Current methods usually formulate the feature representation learning as the CNN parameter θ optimization, and the problem is resolved by minimizing the empirical classification risk of feature representation f on the training set D through backward propagation. The empirical classification risk can be denoted as,

$$\mathcal{J} = \frac{1}{N}\left[\sum_{i=1}^{N} \mathbf{F}\left(\hat{y}_i, y_i\right)\right], \tag{3.2}$$

$$\theta = \arg\min(\mathcal{J}), \tag{3.3}$$

where $\hat{y}_i$ and y_i are respectively the predicted classification score and the target label for the i-th training sample. $\mathbf{F}(.)$ calculates the classification loss for each sample.

In our project, we model this Re-ID issue as 4 sub-tasks under a shared backbone, including global feature learning, classification task for the global feature, classification task for part features, and part transfer for Main-Partial Task pairs. With the shared backbone, these 4 tasks can be jointly trained in our unified feature learning network and achieve the optimal model generalization. Except for global feature learning, the rest 3 tasks are contained within the PMLH. The global feature learning head is introduced in Chapter 3.2.2 and details of the PMLH are included in Chapter 3.2.3.

3.2.2 Global Feature Learning Head

Our part-mentored feature learning network utilizes ResNet-50 as backbone, which is split into two head networks by two separate *res_conv5* residual stages because using

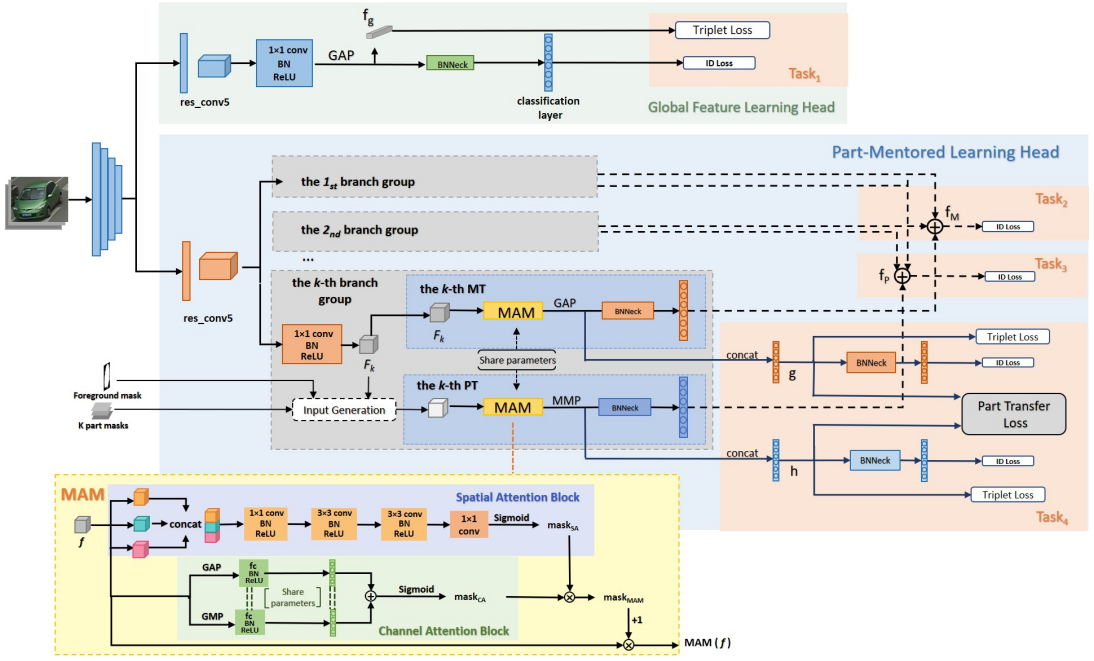

FIGURE 3.5: Architecture of the Part-Mentored Network. The ResNet-50 backbone is split into two sub-networks with two *res_conv5* residual blocks: the Global Feature Learning Head (in green) and the Part-Mentored Learning Head (in blue). As the blue block shows, PMLH is composed of K branches which consists of one Main Task (MT) and one Partial Task (PT). With different inputs, these two tasks are similar in architecture and share one Multi-scale Attention Module (MAM). As shown in the yellow block, MAM utilizes three convolutional layers with multiple dilation ratios to mine multi-scale features in its Spatial Attention Block and obtain a spatial attention mask, $mask_{SA}$. An output channel attention mask, $mask_{CA}$, is generated by Channel Attention Block. Then these two masks are multiplied together to gain $mask_{MAM}$. The orange area respectively illustrate loss functions employed in four sub-tasks. Here $\otimes$ denotes element-wise multiplication and $\oplus$ indicates element-wise summation.

a backbone which shares it might dilute the importance of detailed information [62]. Our global feature learning head aims to conduct a coarse look at the entire image and extract apparent global information to distinguish identities with significant appearance differences. To increase the size of the output feature maps, we remove the last spatial down-sampling operation of both *res_conv5*.

As is shown in Fig.3.5, our global branch feeds the output feature map of the backbone into several convolutional layers, a Global Average Pooling (GAP) and a BNNeck layer [41]. The learning process is supervised by one Triplet loss [19] and one ID loss.

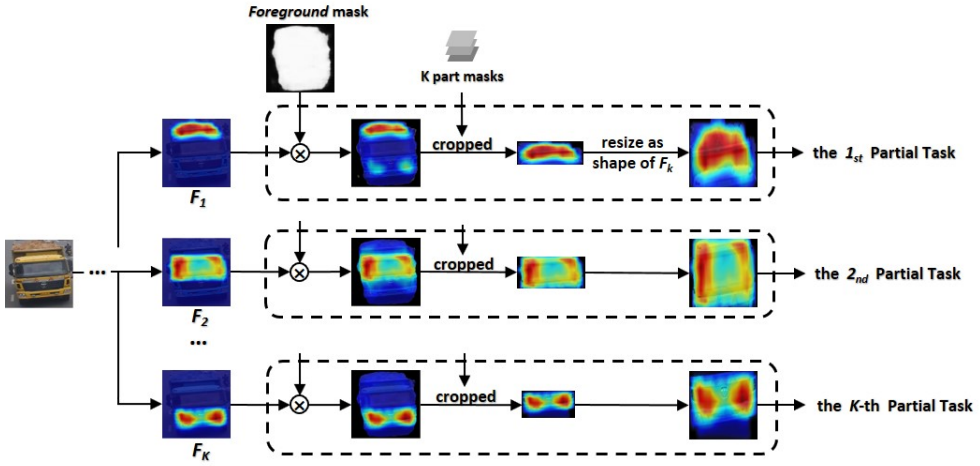

FIGURE 3.6: Input generation for Partial Tasks. The dashed block shows details of the Input Generation module in Fig.3.5. After being multiplied with the refined foreground mask, the feature map F_K is further cropped according to the K-th bounding boxes of the K-th part mask and then resized to the shape of F_K. Here we set $K = 3$ in our experiment and show visualization maps of the three inputs for Part Tasks. It can be observed that these maps respectively focus on three different parts with semantic meaning, i.e., vehicle roof, windscreen and headlights.

3.2.3 Part-Mentored Learning Head

Standard regional feature-based methods [64, 14, 3, 43, 13] simply use a single branch for local feature extraction on each vehicle part, but this shows two weaknesses. First, it requires partial masks for both training and inference. Second, it only learns part-level features within each partial mask. However, partial masks from existing models are not always accurate (see Fig.1.3). Such inaccurate partial masks will introduce noises especially during inference. In our project, we introduce PMLH to perform a second closer look at the localized vehicle parts for more robust, discriminative partial features and aligns them with the global one in a part transfer manner.

Specifically, PMLH is composed of K branches, which respectively concentrate on K vehicle parts localized by the first step, PANet. Each branch utilizes one 1×1 Conv layer to combine relevant channels for the specific vehicle part from the output feature maps of the backbone. After the 1×1 Conv layer, each branch builds two tasks, the Main Task (MT) and the Partial Task (PT), which respectively focuses on the holistic feature space and one vehicle part feature space. All the MTs and PTs are similar in architecture but have different inputs. As Fig.3.5 shows, each MT incorporates one Multi-scale Attention Module (MAM), GAP, one BNNeck layer [41] and a classification layer. The GAP in each MT learns translation and scale-invariant holistic feature. Each

PT replaces GAP with a Mask Max Pooling (MMP) layer, which can be implemented as a Max Pooling layer with pooling size equal to the area of partial masks. Compared with GAP, MMP helps mine more significant local features within the constrained mask region. To handle the two weaknesses of the standard single local branch structure, we don't totally trust part masks. Instead, to learn more robust local features, each part mask branch, i.e., Partial Task, is regarded as a noisy teacher. During training. it guides its student, the correspondent Main Task, to concentrate on the same vehicle part. Such guidance is realized by our proposed sample-wise Part Transfer loss. Since Main Tasks learn features within the entire feature map, F_K, it avoids producing the noises that might be introduced by inaccurate vehicle part masks. Moreover, these Partial Tasks can be removed during inference, which means that our method can also be free from prior part mask generation during testing.

The major difference between each MT-PT pair of one vehicle part lies in their inputs. The input of each MT is the original feature map after the 1×1 Conv layer, F_K, while each PT is equipped with the correspondent partial feature map during training. As illustrated in Fig.3.6, we first multiply F_K with the refined foreground mask. Afterwards, the multiplied feature map is separately cropped according to the bounding boxes of the K part masks from PANet and is then resized as the shape of F_K via bilinear interpolation. In this way, K input feature maps for PTs are obtained. As a result, they can provide guidance for MTs to amplify channels relevant to the correspondent vehicle parts via part transfer, which will be detailed presented in Chapter 3.2.3.2. During inference, our method achieves state-of-the-art performance without PTs, which means our PMLH can be free from prior mask generation of PANet. Results in Chapter 4.1.4 also illustrate that our method evaluated without PTs yields a relatively higher CMC@1. Next, we will introduce details of the key components of our PMLH, i.e., the Multi-Scale Attention module, Part Transfer on MT-PT pairs and Re-ID loss functions for our method.

3.2.3.1 Multi-scale Attention Module

Vehicle components often exhibit large variations in shape, position and scale. For example, the headlights and annual labels are at a small local scale, while the windscreen is at a larger scale. Directly using bottom-to-up single-scale convolution and pooling tends to be ineffective in handling these complex variations. Particularly, as the number of layers increases, some small visual regions will be easily missed in top layers. Therefore, to give an additional soft pixel-level attention refinement within each salient

vehicle part localized by the former hard attention-based PANet, our PMNet constructs a Multi-scale Attention Module (MAM) to handle such scale variation and learn subtle yet distinguished details.

As is depicted in the yellow block in Fig.3.5, MAM transforms the input feature into a spatial attention mask, $\mathbf{mask_{SA}}$, and a channel attention mask, $\mathbf{mask_{CA}}$, respectively from a Spatial Attention Block and a Channel Attention block. Afterwards, these two masks are multiplied together to get a new attention mask, $\mathbf{mask_{MAM}}$. We apply this multi-scale attention mask to the original feature f and the formula can be computed as,

$$\mathbf{MAM}(f) = \mathbf{mask_{SA}} \times \mathbf{mask_{CA}} \times f + f \tag{3.4}$$

In the Spatial Attention Block, to make local regions uniform, we first feed feature f into three 3×3 convolutional (Conv) layers with multiple dilation ratios (1, 2 and 3) and then concatenate the output features. The concatenated feature containing multi-scale information is fed into another 3 Conv blocks to learn spatial attention. In the Channel Attention Block, f is fed into two branches respectively beginning with: Global Average Pooling (GAP) and Global Max Pooling (GMP). Then the two output features are fed into two fully connected layers and added together to gain the $\mathbf{mask_{CA}}$. A Sigmoid function is applied at the end of the Spatial Attention Block and Channel Attention Block to normalize the attention masks.

To learn multi-scale features, some previous works usually use classic Conv filters with multiple kernel sizes (e.g., 3×3, 5×5, 7×7), which overlap with each other at the same output position and thereby causes much redundant information. However, as Fig.3.7 shows, with convolutional layers of multiple dilation ratios, only the center position of a feature representation is overlapped. Therefore, our MAM reduces redundancy with different receptive fields.

Note that the self-attention in PANet and the Multi-scale Attention Module make up a combination of hard part-level and soft pixel-level attention. Such a two-stage attention method leverages the second stage module to further refine the first stage results and conducts a coarse-to-fine multi-grained search in vehicle Re-ID compared with common one-stage attention methods [60, 59].

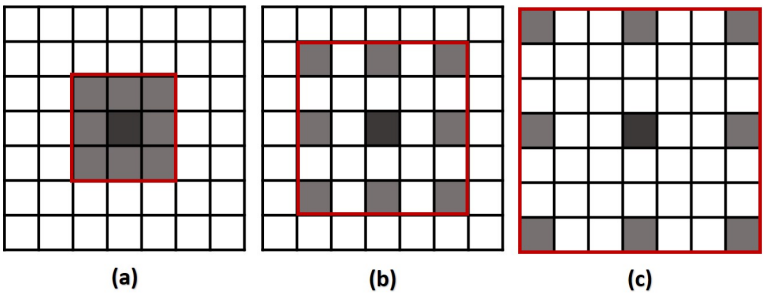

FIGURE 3.7: Examples of dilated convolution and common convolution for the same feature map. In (a)(b)(c), the red rectangles respectively show convolution filters of 3×3, 5×5 and 7×7 while the grey areas illustrates the convolutional layers with different dilation ratio of 1,2 and 3.

Remarks Our hard-soft attention selection is similar to some existing works[13, 31] but differs in design: $(\mathbf{1})$ Compared with TAMR [13], our approach optimizes Re-ID performance without extra supervision during hard attention and exploits multi-scale features with simple dilated layers during soft attention. $(\mathbf{2})$ Compared with HA-CNN[31], our method learns local features on K different vehicle parts, which reduces information redundancy caused by similar part masks predictions.

3.2.3.2 Feature Alignment via Part Transfer

Our PMLH leverages each PT, the noisy teacher, to mentor the learning of its correspondent student, MT, so that they can both select and combine the part-relevant channels via the introduction of inductive bias. This means that built on the same backbone, PTs learn the part-specific information to construct the part-level representations for fine-grained classification and transfer their concept to MTs via hard parameter sharing of MAM as well as a Part Transfer loss.

The Part Transfer loss is demonstrated in the orange area of Task_4 in Fig.3.5. We firstly concatenate the K features from the K MTs and PTs to make up feature g and h, respectively. Then after normalizing g and h of each sample, we compute the mean distances between them in a batch and penalize those beyond our pre-defined margin, γ. The formula can be written as,

$$\mathcal{L}_{\mathbf{PT}} = \max\left\{\frac{1}{B}\left[\sum_{i=1}^{B}\left\|\mathcal{N}(g_i) - \mathcal{N}(h_i)\right\|_2\right] - \gamma, 0\right\}, \tag{3.5}$$

where $\mathcal{N}(.)$ indicates the normalization function and B is the number of samples in a batch. g_i and h_i respectively denote the concatenated feature g and h for the i-th sample. In our experiments, we set γ as 0.05. This loss encourages Main Tasks and Partial Tasks to pull each other closer within the γ distance, which makes sure that guided by local information of Partial Tasks, each Main Task focuses on the specific vehicle part.

Remarks PMNet contributes in the following aspects: $(\mathbf{1})$ It compensates inaccurate part masks generated by the former weakly-supervised localization step, thereby improving overall Re-ID performance. This is an aspect often overlooked by current vehicle Re-ID community. $(\mathbf{2})$ PMNet frees the model from prior part-level constraints during inference, which further saves time and computation cost.

3.3 Multi-task Learning for Vehicle Re-ID

As is depicted in the orange areas in Fig.3.5, our model divides this Re-ID issue into 4 sub-tasks, Task$_j$, $j \in [1, 4]$ under a shared backbone. These 4 sub-tasks are respectively global feature learning, ID classification sub-task for global feature, ID classification sub-task for part features and part transfer sub-task. By sharing experience with each other, these four sub-tasks improve model generalization and lead to better prediction performance.

Task$_1$: the global feature learning sub-task is supervised by one Cross Entropy loss and one Triplet loss [19] on its output global feature, f_g. For the following sub-tasks, we combine both feature summation and feature concatenation in our feature alignment to enable effective feature re-usage and re-exploration [6]. Task$_2$ and Task$_3$ are two ID classification sub-tasks, the K features from the classification layer of MTs, f_M^i, $i \in [1, K]$, and PTs, f_P^i, $i \in [1, K]$, are respectively summed up to get the final classification features, f_M and f_P. The formula can be written as,

$$\begin{aligned} f_M &= \sum_{i=1}^{K} f_M^i \\ f_P &= \sum_{i=1}^{K} f_P^i \end{aligned} \tag{3.6}$$

Cross Entropy loss is adopted on f_M and f_P in Task$_2$ and Task$_3$, respectively. As regards Task$_4$, feature concatenation is applied in this part transfer sub-task to obtain feature

representation for Main Tasks (MT), g and feature representation for Partial Tasks (PT), h. Afterwards, we align the global-local feature spaces by the part transfer sub-task with one Part Transfer loss, $\mathcal{L}_{\mathbf{PT}}$, one Cross Entropy loss on g, one Cross Entropy loss on h and two Triplet losses respectively on g and h.

In total, there are five ID classification losses respectively on f_g, f_M, f_P, g and h, which can be linearly summed up with 5 weights, α_j, $j \in [1, 5]$. The formula can be written as,

$$\mathcal{J}_{\mathbf{ID}} = \sum_{j=1}^{5} \left[\alpha_j \mathcal{L}^{\mathbf{j}} \left(\hat{y}^j, y^j \right) \right], \tag{3.7}$$

where $\mathcal{L}^{\mathbf{j}}(.)$ denotes the Cross Entropy loss on the j-th classification score and α_j indicates the weight for the j-th ID loss. $\hat{y}_i^j$ and y_i^j are respectively the predicted classification score and target label of the i-th training sample for the j-th ID loss.

Obviously in Equation (3.7), it is also essential to carefully select the weighting of each ID loss, α_j, for jointly balanced learning of all the five ID losses under a common backbone. Hence, instead of using a naïve linear sum of multiple objective losses, we leverage Homoscedastic Uncertainty Learning (HUL) [24] to automatically learn the optimal weights without extra grid search. HUL introduces a noise parameter σ_j for each variable $\hat{y}^j$. As the noise σ_j increases, the loss weight, α_j, for its respective objective decreases. Assume that the classification likelihood is adapted using a softmax loss function with a noise scalar σ_j to compress the scaled version of the model output :

$$p\left(\hat{\mathbf{y}}^{\mathbf{j}} \mid \mathbf{f}^{\mathbf{W^j}}(\mathbf{x}), \sigma_j \right) = \text{Softmax}\left(\frac{1}{\sigma_j^2} \mathbf{f}^{\mathbf{W^j}}(\mathbf{x}) \right) \tag{3.8}$$

Hence, the minimization objective of this HUL-based multi-task ID loss, i.e., Equation (3.7), can be converted to,

$$\mathcal{J}_{\mathbf{ID}} \approx \sum_{j=1}^{5} \frac{1}{\sigma_j^2} \mathcal{L}^{\mathbf{j}} \left(\hat{y}^j, y^j \right) + \sum_{j=1}^{5} \log \sigma_j \tag{3.9}$$

In total, the five ID losses, our proposed sample-wise Part Transfer loss and Triplet loss [19] are combined to supervise our multi-task learning during training. Hence, Equation (3.2) and the overall objective, $\mathcal{J}$, can be reformulated as,

$$\mathcal{J} = \mathcal{J}_{\mathbf{ID}} + \mathcal{L}_{\mathbf{PT}} + \mathcal{L}_{\mathbf{Triplet}}. \tag{3.10}$$

Chapter 4

Experiments

In this chapter, we first introduce the two widely utilized vehicle Re-ID datasets and evaluation metrics. Then we compare our method with the state-of-the-arts and conduct ablation studies to validate the effectiveness of each component.

4.1 Datasets and Evaluation Metrics

4.1.1 Datasets

As is shown in Table.4.1, we use the two datasets to evaluate the performance of our method.

VeRi776 [39] consists of around 51,035 images from 776 vehicles, including 576 identities for training and 200 for testing. The standard query and gallery sets respectively contain 1,678 and 11,579 images. Images in this dataset are captured by 20 cameras under different viewpoints, making it one of the most challenging vehicle Re-ID datasets. Additionally, it also provides other vehicle information, including model, color and trajectory clues.

VehicleID [35] contains totally 221,763 images for about 26,267 vehicles. The test set is divided into three sizes (small, medium and large). During testing, we randomly select one vehicle image as the gallery image and regard the rest as the query set. Images of this dataset are either captured under front or rear view.

TABLE 4.1: Detailed information of the vehicle Re-ID datasets that we utilize.

Splits		VeRi776	VehicleID		
Train	Identities	576	13164		
	Images	37746	113346		
	Cameras	20	/		
Query	Identities	200	Small	Medium	Large
			800	1600	2400
	Images	1678	800	1600	2400
	Cameras	19	/		
Gallery	Identities	200	800	1600	2400
	Images	11579	5693	11777	17377
	Cameras	19	/		

4.1.2 Evaluation Metrics

We evaluate the model accuracy with two types of metrics in Re-ID: mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC). In specific, mAP computes the averaged area under the Precision-Recall curve (AP) for all the query images. CMC@1, 5 respectively measures the probability to locate at least one true positive in the top-1, 5 ranks.

4.1.3 Implementation Details

4.1.3.1 Training

Implemented with PyTorch, our model is based on the well-known Bag-of-Tricks Re-ID baseline[1] [41]. We train PANet for 100 epochs on VeRi776. The batch size is 64 and the learning rate is 1×10^{-4}. In our PMNet, we randomly select P identities and Q images per vehicle to construct a training batch. Finally, the batch size, B, is computed as $B = P \times Q$. Here, we set $P = 4$, $Q = 8$, $B = 32$. It should also be noted that the base learning rate of PMNet is 1.5×10^{-4} with warm-up strategy and images are resized to 256×256. For data augmentation, we apply Random Erasing Augmentation and random horizontal flip with a probability of 0.5. As regards other hyper-parameters, in our experiments, the number of vehicle parts localized by PANet, K, is set as 3; the margin for Triplet loss [19], β, is set as 0.7. In addition, we use Adam as the optimizer.

[1] https://github.com/DTennant/reid_baseline_with_syncbn

Since there is no special dataset for occluded vehicle Re-ID, we simulate the occlusion by using a simple data augmentation technique with a probability of 0.3. For each training batch, we first randomly select an area as the obstruction region. These areas are then filled with random color patches.

4.1.3.2 Inference

To evaluate our method, we obtain g, h and global feature, f_g respectively after the BNNeck layers of our PMLH and global feature learning head. Then we calculate the Cosine Distances, D_g, D_h and D_{f_g}. Final feature distance between a probe and gallery set is calculated as $\lambda_1 D_{f_g} + \lambda_2 D_g + \lambda_3 D_h$, where λ_i, $i \in [1, 3]$ denotes the optimal weights learned by HUL. For instance, in our experiment of evaluating PMANet on VeRi776, $\lambda_1 = 5.33$, $\lambda_2 = 4.43$, $\lambda_3 = 4.25$. In our PMANet^o in Table.4.2 and Table.4.3, the Partial Tasks (PTs) are removed during evaluation, which means that λ_3 is set to 0.

FIGURE 4.1: Visualization of vehicle Re-ID ranking list on VeRi776. The images on the first column are the query images and the rest show retrieved gallery images. For each query sample, the first and second row respectively shows the top-5 results by baseline and PMANet. we draw the correct and false matched vehicle images respectively with green and red rectangles.

TABLE 4.2: Performance (%) comparisons with the state-of-the-arts on VeRi776. EA is short for extra annotation labeled by humans.

Method	Size	EA	mAP	CMC@1	CMC@5
SPAN[3]	256×256	Y	68.9	94.0	97.6
PRReID[14]	256×256	Y	70.2	92.2	97.9
MVAN[59]	256×256	Y	72.5	92.6	97.9
CFVMNet[57]	256×256	Y	77.1	95.3	98.4
PVEN[43]	256×256	Y	79.5	95.6	98.4
RAM[37]	224×224	N	61.5	88.6	94
SAN[47]	256×256	N	72.5	93.3	97.1
SAVER[26]	256×256	N	79.6	96.4	98.6
Baseline[41]	256×256	N	77.2	95.7	97.9
Our PMANet^o	256×256	N	81.7	**96.7**	**98.6**
Our PMANet	256×256	N	**81.8**	96.4	**98.6**
CFVMNet[57]+RR	256×256	Y	81.5	94.8	96.6
Our PMANet+RR	256×256	N	**89.8**	**97.3**	**98.2**

4.1.4 Comparison with the State-of-the-Art Methods

We first compare the performance of PMANet with a variety of recent approaches on VeRi776 [39] and VehicleID [35]. Experimental results are respectively tabulated in Table.4.2 and 4.3, where EA is short for extra annotation manually labeled.

4.1.4.1 Experiments on VeRi776

We adopt CMC@1, CMC@5 and mAP as the evaluation protocol on VeRi776. As Table.4.2 shows, in terms of whether using extra annotations, existing local-based vision methods can be divided into two categories. [14, 3, 43, 59, 57] rely on expensive key-point labels or part annotations, while [37, 47, 26] are trained without extra manual labels. [37, 47, 57] also incorporate other vehicle attributes (e.g., vehicle model, color) to enhance feature expressions. Among the EA-based works [14, 3, 43, 59, 57], PVEN [43] performs best in all the three metrics while SAVER [26] achieves relatively high results in the Non-EA category. The results in Table.4.2 demonstrate that our PMANet sets a new state-of-the-art, surpassing PVEN and SAVER respectively by 2.3%, 2.2% in mAP. Besides, by removing PTs during inference, PMANet^o yields an even higher CMC@1. In our PMANet+RR version, we adopt [84], the re-ranking method, as a Re-ID post-processing technique. As a result, we can observe that the mAP greatly rises to

89.8% and the CMC@1 increases to 97.3%. Fig.4.1 demonstrates some sample vehicle Re-ID results on VeRi776. Compared with the retrieved ranking list of the baseline, our method clearly produces more reliable results. These top-5 results show that our method is more robust to viewpoint variation, low resolution, background clutter and is also more capable of mining fine-grained local clues to distinguish near-identical vehicles.

TABLE 4.3: Performance (%) comparisons with the state-of-the-arts on VehicleID. EA is short for extra annotation labeled by humans.

Method	Size	EA	Small		Medium		Large	
			CMC@1	CMC@5	CMC@1	CMC@5	CMC@1	CMC@5
TAMR[13]	256×256	Y	66.0	79.7	62.9	76.8	59.7	73.9
PRReID[14]	256×256	Y	78.4	92.6	75	88.3	74.2	86.4
CFVMNet[57]	256×256	Y	81.4	94.1	77.3	90.4	74.7	88.7
PVEN[43]	256×256	Y	**84.7**	**97.0**	**80.6**	**94.5**	**77.8**	**92.0**
RAM[37]	224×224	N	75.2	91.5	72.3	87.0	67.7	84.5
SAN[47]	256×256	N	79.7	94.3	**78.4**	**91.3**	**75.6**	**88.3**
SAVER[26]	256×256	N	**79.9**	**95.2**	77.6	91.1	75.3	88.3
Our PMANet^o	256×256	N	85.0	**97.5**	**79.8**	**94.1**	**76.9**	91.6
Our PMANet	256×256	N	**85.3**	97.4	79.8	94.0	76.7	**91.7**

4.1.4.2 Experiments on VehicleID

Since there is only one true match for each query vehicle in the gallery set in VehicleID, merely CMC@1 and CMC@5 are compared in this dataset. Table.4.3 illustrates the comparison results on the small, medium, and large test sets. Depending on the usage of extra annotation, Table.4.3 groups existing methods into two categories. Results show that our proposed PMANet beats all the models in the Non-EA category, with an average improvement of 2.6% in CMC@1 and 2.8% in CMC@5. In the EA-based category, with an increase of 0.6% in CMC@1 and 0.4% in CMC@5, our method performs best on the small test set. As for the medium and large test sets, our method produces the second-best CMC@1 and CMC@5. Although our results are slightly lower than those of PVEN [43] on the medium and large test set, our method saves the cost for dense key-point labels required by PVEN. Our approach is more suitable for practical applications.

TABLE 4.4: Ablation studies about each component of our PMANet on VeRi776. ✓ in each row denotes the modules that are included in this experiment. Exp-7 denotes our entire method, PMANet.

Experiment Number	Global Feature Learning head	PANet	PMLH	MAM	mAP	CMC@1	CMC@5
Exp-1	✓				72.4	94.6	97.2
Exp-2	✓	✓	$\mathcal{R}_2$		78.5	95.7	98.0
Exp-3	✓	✓	✓		79.5	96.4	98.1
Exp-4	✓	$\mathcal{R}_1$	✓	✓	79.9	96.4	98.3
Exp-5		✓	✓	✓	78.1	95.3	97.8
Exp-6	✓	✓	$\mathcal{R}_2$	✓	78.2	95.6	98.0
Exp-7	✓	✓	✓	✓	**81.8**	**96.4**	**98.6**

4.1.5 Ablation Study

4.1.5.1 Component Analysis

In this section, we conduct comparative experiments to validate the effectiveness of the proposed components, including Global Feature Learning Head (GFLH), Part Attention Network (PANet), Part-Mentored Learning Head (PMLH) and Multi-scale Attention Module (MAM). Detailed results are tabulated in Table.4.4 and all the experiments are evaluated with Partial Tasks. In Table.4.4, $\mathcal{R}_1$ indicates that in Exp-4, PANet is replaced by Uniform Division for part mask generation and $\mathcal{R}_2$ denotes that PMLH is replaced by K single branches with plain convolution layers for local feature learning.

Global Feature Learning Head. By comparing Exp-5 and Exp-7, we can observe a dramatic increase of 3.7%, 1.1% and 0.6% in mAP, CMC@1 and CMC@5, respectively. This validates the effectiveness of our Global Feature Learning Head and shows that this global head and PMLH provide complementary information for each other.

Part Attention Network. Since PMLH requires K prior partial masks to conduct part-mentored feature aggregation, we replace PANet with a simple Uniform Division for mask generation in Exp-4. Uniform Division, which evenly splits the feature map into K stripes, is a common way utilized in [37, 47]. As is tabulated in Table.4.4, PMANet beats Exp-4 by 1.9% in mAP and 0.3% in CMC@5. Moreover, such a naive splitting strategy is unstable. As Fig.4.2 depicts, especially when a car body is unevenly

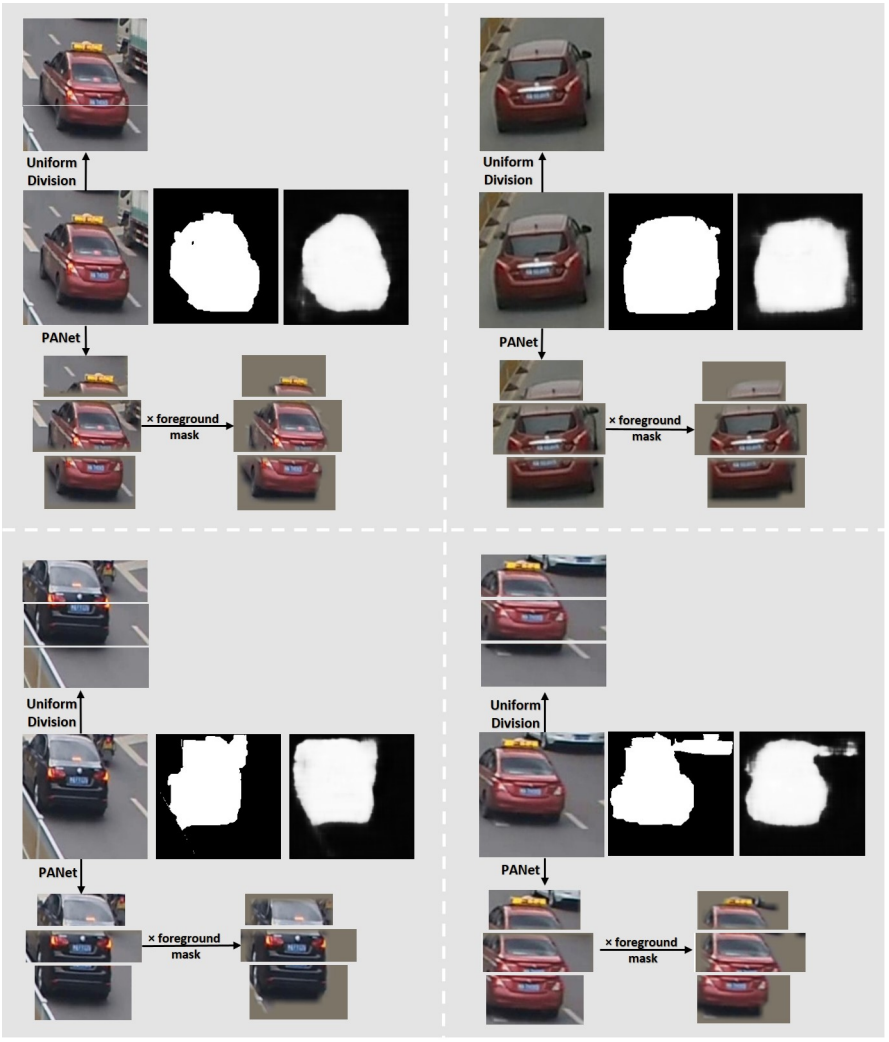

FIGURE 4.2: Examples of generated masks by Uniform Division and our PANet. The sub-figure on the first row illustrate three parts vertically split by Uniform Division. The three figures on the second row respectively represent the sample image, the coarse foreground mask by GrabCut [52] and the refined one by our PANet. On the third row, the two sub-figures exhibit the K different part masks predicted by PANet and these K part masks multiplied with the refined foreground mask.

distributed in the image, Uniform Division suffers from spatial misalignment and might miss some crucial information. In contrast, our PANet is able to locate different salient vehicle parts (e.g., vehicle roof, windscreen, lights) with almost all the subtle cues, such as personalized decorations and inspection marks. It is also robust to background clutter and noises.

Part-Mentored Learning Head. Instead of using part-guided learning structure, Exp-2 and Exp-6 replace PMLH with K plain convolutional branches respectively with

and without MAM for detailed local feature extraction. Comparing Exp-6 with Exp-7, PMLH boosts the performance by 3.6% in mAP and 0.8% in CMC@1. Comparing Exp-2 with Exp-7, our PMANet also leads to an increase of 3.3% in mAP and 0.7% in CMC@1. These two comparisons prove that our part transfer manner shows a superior capability of extracting part-level clues than plain convolution branches.

Multi-scale Attention Module. The next step is to investigate the effectiveness of our soft attention, Multi-scale Attention Module (MAM). PANet automatically localizes prominent regions but treats pixels within each region equally. Afterwards, MAM performs a second closer look at these parts for pixel-wise local feature enhancement to decrease inter-class similarity. Comparing Exp-7 with Exp-3, MAM brings an improvement of 2.3% in mAP and 0.5% in CMC@5. Furthermore, we visualize response maps on K branches of PMLH to explore how our MAM affects feature learning and the visualized maps are shown in Fig.4.3. By comparing response maps before and after the MAM in each Main Task as well as Partial Task, we can observe that the MAM assists in capturing and amplifying more subtle clues. In addition, we also compare MAM with other attention methods to verify its superiority. Among existing approaches in vehicle ReID, SE-Net [21] is a channel-wise attention commonly utilized in vision models; TAMR [13] leverages Residual Attention Module in its regional feature learning. Therefore, we replace MAM with these two attention modules on the second and third row of Table.4.5. Clearly, mAP drops by 0.8% and 1.1%, respectively. Therefore, compared with SE-Net and Residual Attention Module employed in TAMR [13], our MAM exhibits better ability in handling complex scale variation and mining multi-grained clues.

TABLE 4.5: Comparison experiments to validate Multi-scale Attention Module (MAM) on VeRi776. For the first and second row, MAM in PMNet is repaced by SE-Net [21] and Residual Attention Module adapted in [13]

Method	mAP	CMC@1	CMC@5
PANet+SE-Net	81.0	**96.8**	98.2
PANet+Residual Attention Module	80.7	96.7	**98.7**
PMANet	**81.8**	96.4	98.6

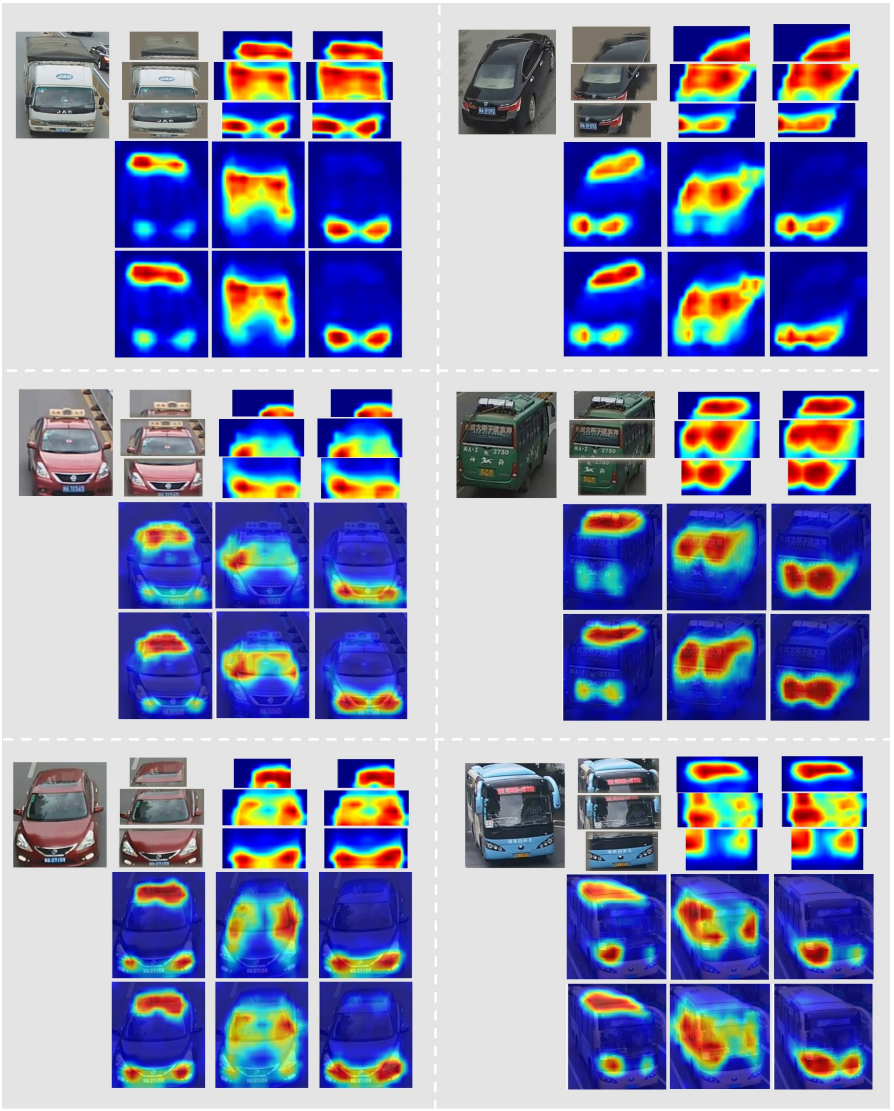

FIGURE 4.3: Examples of response maps. the first row of each sub-figure respectively shows the original image, K part masks by PANet, response maps of K Partial Tasks before MAM, and response maps of K Partial Tasks after MAM. The second and the third row of each sub-figure show each response map of the K Main Tasks before and after MAM, respectively.

4.1.5.2 Validation of Part Transfer

Instead of learning features with plain Conv layers in a single branch [13, 3, 43], our Part-Mentored learning Head (PMLH) introduces K Main-Partial Task pairs and leverages hard parameter sharing along with a sample-wise Part Transfer loss to convey part information from Partial Tasks (PT) to Main Tasks (MT). Part transfer aims to align the feature space of each MT-PT pair. As is described in Sec.4.1.5.1, PMLH exhibits a superior local feature learning ability compared with using plain convolutional branches.

TABLE 4.6: Experiments to validate the effectiveness of part transfer, which is realized via parameter sharing and our Part Transfer loss, $\mathcal{L}_{\mathbf{PT}}$, on VeRi776. *w/o* denotes without.

Method	mAP	CMC@1	CMC@5
w/o parameter sharing	81.1	96.3	98.4
w/o $\mathcal{L}_{\mathbf{PT}}$	80.2	**96.7**	**98.6**
w/o parameter sharing & $\mathcal{L}_{\mathbf{PT}}$	80.6	96.4	98.3
PMANet	**81.8**	96.4	**98.6**

Moreover, Fig.4.3 shows visualized response maps of PTs and MTs in each branch. Apparently, response maps of each MT-PT pair mainly concentrate on the same vehicle part, and the K branches concentrate on K different regions which are located by PANet. This means that realized by parameter sharing and the novel Part Transfer loss, part transfer successfully guides the learning of MTs with concept conveyed in PTs. Next, we conduct experiments to verify the function of parameter sharing and our Part Transfer loss, $\mathcal{L}_{\mathbf{PT}}$, in Table.4.6. With a respective increase of 1.6%, 0.7% in mAP, $\mathcal{L}_{\mathbf{PT}}$ and parameter sharing prove to be beneficial.

TABLE 4.7: Experiments to investigate Homoscedastic Uncertainty Learning on VeRi776. *WR* denotes weighting ratio.

Method	mAP	CMC@1	CMC@5
w/o HUL (*WR*:1:1:1:1:1)	80.5	96.5	98.0
w/o HUL (*WR*:2:1:1:1:1)	80.4	96.2	**98.7**
w/o HUL (*WR*:4:1:1:1:1)	79.8	96.4	98.0
Add HUL from Epoch30	80.7	96.1	98.1
PMANet	**81.8**	**96.4**	98.6

4.1.5.3 Validation of Our Loss Functions

In this sub-section, we respectively present our experiments to investigate multi-task learning with Homoscedastic Uncertainty Learning and our combination of feature summation as well as concatenation.

Multi-task Learning with Homoscedastic Uncertainty Learning. In our project, we model this Re-ID issue as 4 sub-tasks. With the shared backbone, these 4 different tasks can be trained end-to-end in our unified feature learning network, PMNet, and achieve the optimal model generalization with the help of Homoscedastic Uncertainty Learning (HUL). Here we first validate the effectiveness of our multi-task learning idea in Table.4.8 and then investigate the superiority of adopting HUL in Table.4.7 on VeRi776. All the experiments are evaluated with PTs. In Table.4.8, *w/o* classification tasks means to remove the identity classification sub-task for global features and identity classification sub-task for part features while the second row denotes an experiment without the part transfer sub-task. In comparison with results on the first two rows in Table.4.8, PMANet performs best. In detail, compared with the first row, our PMANet increases the mAP, CMC@1 by 3.4% and 0.7%. In comparison with results on the second row, it is observed that our part transfer sub-task boosts the mAP by 2.7%. This demonstrates that our 4 different sub-tasks provide complementary information for each other, which helps increase inter-instance variance and thus facilitate our final fine-grained re-identification task for vehicles, especially for near-duplicated identities.

For the first three rows in Table.4.7, we remove HUL and manually select three different weights. In the fourth row, we incorporate HUL at training epoch 30 to relieve the influence of model noises on weight learning at the beginning of training. This is because the model is quite unstable and sensitive to noises at the beginning of training, thus making the noise parameters introduced by HUL, σ_j, increase rapidly. Since the ID loss weights are inversely proportional to the noise parameters, they will decline with the increase of σ_j, making it hard for the model to converge. Apart from speeding the convergence process during training, results in Table.4.7 illustrate that HUL helps save the time cost by extra manual tuning and is robust to hyper-parameter changes.

TABLE 4.8: Experiments to verify the superiority of our multi-task learning with four sub-tasks and combination of feature summation & concatenation on VeRi776.

Method	mAP	CMC@1	CMC@5
w/o classification tasks	78.4	95.7	97.9
w/o part transfer task	79.1	95.6	98.1
w/o feature summation	79.8	96.3	98.2
w/o feature concatenation	80.6	96.0	98.2
PMANet	**81.8**	**96.4**	**98.6**

Combination of Feature Summation and Concatenation. As regards feature alignment, we incorporate feature summation with concatenation to enable both feature reusage and re-exploration. Statistics in the third and fourth row in Table.4.8 show that our combined alignment structure achieves the optimal result in terms of mAP, CMC@1 and CMC@5. It is also observed that our model merely with feature concatenation is more effective than that merely with feature summation, yielding a 0.8% increase at mAP.

4.1.6 Comparisons with Variants of PMANet

In this section, we conduct comparative experiments on VeRi776 to explore the influence of the variants of our PMANet.

TABLE 4.9: Experiments to validate the effectiveness of splitting the trunk branch with two *res_conv5* on VeRi776.

Method	mAP	CMC@1	CMC@5
Share *res_conv5*	79.4	96.1	98.0
PMANet	**81.8**	**96.4**	**98.6**

4.1.6.1 Effectiveness of splitting the trunk branch with two *res_conv5* residual blocks

Our Part-Mentored Network utilizes ResNet-50 as backbone, which is divided into two sub-networks by two *res_conv5* residual blocks: global feature learning head and the Part-Guided Learning Head. MGN [62] explains that utilizing the entire backbone with shared *res_conv5* might dilute the importance of local clues because the last residual stage of ResNet-50 mainly responds to different levels of detailed information on images. We conduct experiments respectively with shared *res_conv5* and separate *res_conv5* in Table.4.9, and results show that adopting separate *res_conv5* achieves an improvement of 2.4% in mAP, 0.3% in CMC@1 and 0.6% in CMC@5. This also proves that these two sub-networks learn complementary information.

TABLE 4.10: Performance (%) comparisons of our proposed method with different values of K on VeRi776. In PMANet^o, all the Partial Tasks are removed during inference.

Method	Variant K	mAP	CMC@1	CMC@5
PMANet^o	$K=2$	80.7	96.0	98.1
	$K=3$	**81.7**	96.7	**98.6**
	$K=4$	**81.7**	**96.9**	98.1
PMANet	$K=2$	79.8	96.0	97.7
	$K=3$	**81.8**	**96.4**	**98.6**
	$K=4$	81.2	**96.4**	98.2

4.1.6.2 Comparisons with variant K

The next step is to compare the performance of our proposed method with different values of variant K. As is tabulated in Table.4.10, we conduct comparison experiments when K equals to 2, 3 and 4, respectively, on VeRi776. Clearly, our PMANet and PMANet^o performs the best when $K=3$, yielding an average increase of 1.3%, 0.5% in mAP and 0.65%, 0.5% in CMC@5. Additionally, the performances rank second when $K=4$.

TABLE 4.11: Comparison experiments in occluded vehicle Re-ID. For the small test set, we use the same test set designed by [23]. With 1678 identities for query, the large test set is constructed based on the test set of VeRi776 [39]

Method	Small			Large		
	mAP	CMC@1	CMC@5	mAP	CMC@1	CMC@5
ASAN[23]	53.3	68.8	90.1	-	-	-
Baseline	60.1	70.7	90.7	54.5	75.9	85.8
PMANet	85.3	**96.0**	**99.7**	65.7	85.3	91.6
PMANet^o	**85.7**	**96.0**	**99.7**	**66.2**	**86.0**	**91.8**

4.1.7 Occluded Vehicle Re-ID

To validate the robustness to occlusion, we firstly show some visualization results of the baseline and our PMANet in Fig.4.4. We can observe that the baseline often wrongly regards the occlusion color as the texture of the query image, while the foreground masks generated by PANet successfully avoid simulated occlusion color blocks. Therefore, our

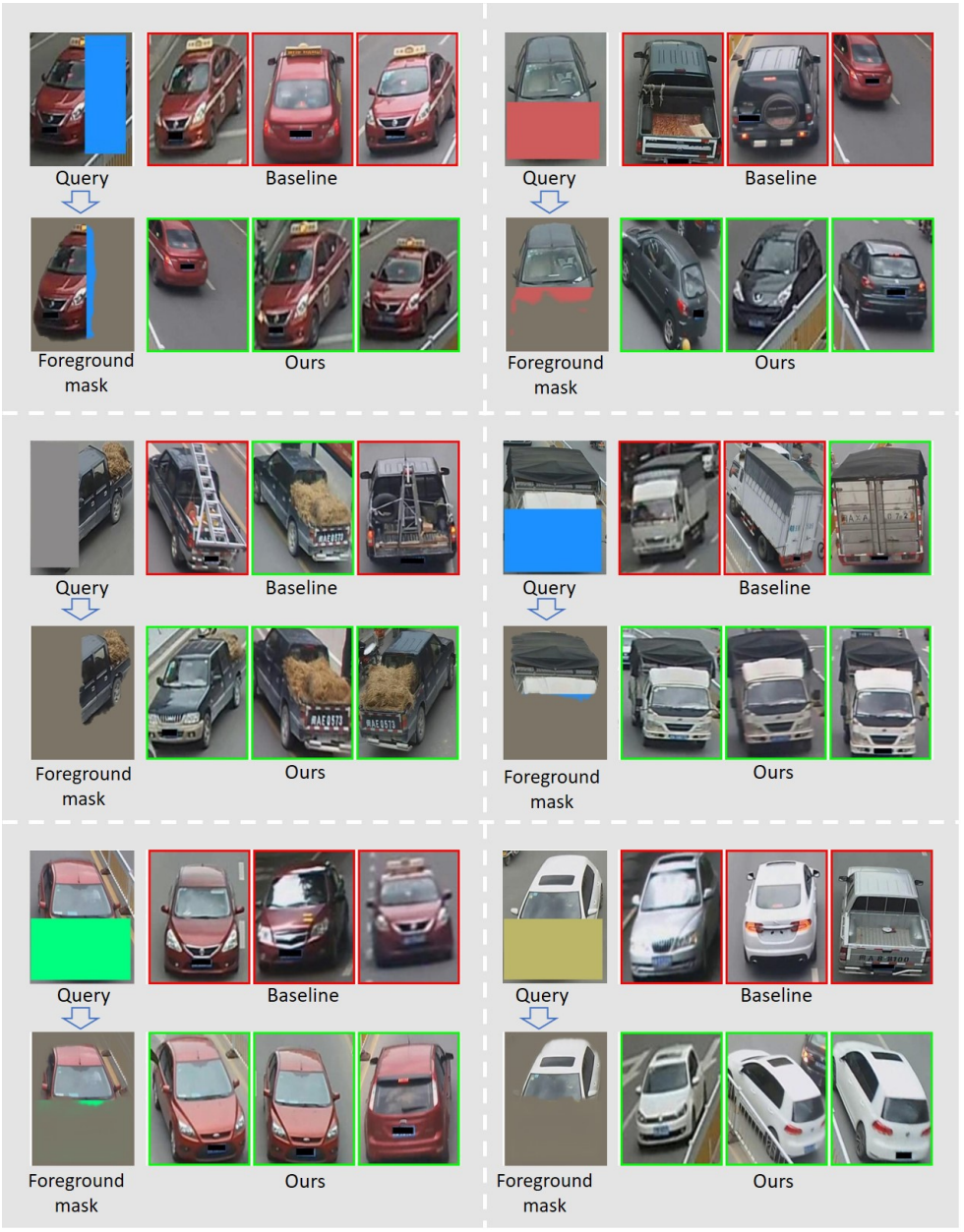

FIGURE 4.4: Visualization results on occluded test set by [23]. In each sub-figure, images on the first column exhibit the query and the foreground mask generated by the first step of PMANet. Images on the second column show the top-3 retrieval results of the baseline and our method. Green and red rectangles respectively denote correct and false results.

method effectively achieves robustness to viewpoint variation under occluded scenarios. To give a fair comparison, we compare our method with the baseline and the state-of-art ASAN [23] on the small test set designed by [23], as shown in Table.4.11. This small test set totally includes 60 identities and 600 images, with 300 for query and 300 for gallery. Our PMANet surpasses the others with a significant improvement of over 25%

in both mAP and CMC@1 on this small test set. Besides, we construct a large test set by applying random color patches to the query set of VeRi776 as shown in Fig.4.4. This large test set contains 200 identities and 11579 images, with 1678 for query and the rest of the images for gallery. Compared with the baseline, our method still performs best by over 10% in mAP, verifying the generalization of our model.

Chapter 5

Conclusion and Recommendations

In this chapter, we make a conclusion for this thesis. Then, we describe current challenges of our proposed method and provide some potential solutions for further research in Chapter 5.2.

5.1 Conclusion

This thesis first presents existing challenges, motivation, and literature review for vehicle Re-ID. Afterwards, we propose a Part-Mentored Attention Network (PMANet) consisting of two-stage attention selection, a weakly-supervised Part Attention Network (PANet) for vehicle part localization and a Part-Mentored Network (PMNet) for feature aggregation. PANet, which is simple in structure and easy to optimize, locates informative vehicle parts under weak supervision with part-level hard attention. Then PMNet employs soft pixel-level attention as the second stage refinement to learn more detailed multi-scale clues within each predicted vehicle part. To address the two weaknesses of the standard single local feature learning branch structure, PMNet builds one Main task and Partial Task to transfer the concept learned in each vehicle part. Each Partial Task is regarded as a noisy teacher to guide the learning of Main Task for more robust local features. This Re-ID issue is modeled as 4 sub-tasks under a shared backbone and experiments demonstrate that our method beats existing approaches with an average increase of 2.63% in CMC@1 on the large dataset, VehicleID [35], and 2.2% in mAP on the small dataset, VeRi776 [39], without any additional annotations. Moreover, results

on occluded vehicle Re-ID test set show the robustness to background interference and occlusion.

5.2 Recommendations for further research

5.2.1 Under-fitting and Over-fitting

As is demonstrated in Chapter 4, the number of images in VehicleID [35] is 4 times that of VeRi776 [39], which requires the models to be equipped with good generalization capabilities. Therefore, existing methods often overfit on small datasets but underfit on large ones. To address this data sparsity issue, one possible solution is to introduce a network which transfers generalizable feature representations learned from large Re-ID datasets to small datasets. In specific, we can design a local-based feature learning network with fewer parameters for smaller datasets. Our method learns detailed discriminative clues and helps guide the training of this smaller network in a teacher-student manner.

5.2.2 Viewpoint Variation

As is mentioned in Chapter 1, two major challenges in vehicle Re-ID lie in the minor inter-instance variance caused by similar identities and the large intra-instance discrepancy caused by different viewpoints. Our PMANet handles the near-duplicated challenge with part-mentored feature alignment, and the detailed discriminative clues learnt addresses the viewpoint variation problem to certain extent. However, our method is still sensitive inference, such as extreme viewpoint changes and low resolution. Here, we point out two potential approaches to cope with these issues, respectively.

Incorporation of view-aware information. One possible solution is to learn view-aware information in our global feature learning head. Specifically, we can split the images into two latent groups, Same-view and Different-view. Afterwards, view-aware constraints can be introduced to pull the samples with a same ID but different viewpoints close to each other. In this way, this global head fuses viewpoint-relevant information into feature embedding. Nevertheless, this approach poses a potential challenge: how to

coarsely divide vehicle images into the Same-view and Different-view groups without additional labels.

Reconstruction of high-resolution images. We can adapt our PANet to reconstruct a new high-resolution vehicle image using the input sample in this solution. Then, instead of the original image, the output image is fed into our PMNet for global-local feature aggregation

Bibliography

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. "Learning to see by moving". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 37–45.

[2] Yan Bai et al. "Group-sensitive triplet embedding for vehicle reidentification". In: *IEEE Transactions on Multimedia* 20.9 (2018), pp. 2385–2399.

[3] Tsai-Shien Chen et al. "Orientation-aware vehicle re-identification with semantics-guided part attention network". In: *European Conference on Computer Vision*. Springer. 2020, pp. 330–346.

[4] Weihua Chen et al. "Beyond triplet loss: a deep quadruplet network for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 403–412.

[5] Weihua Chen et al. "Beyond triplet loss: a deep quadruplet network for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 403–412.

[6] Yunpeng Chen et al. "Dual path networks". In: *arXiv preprint arXiv:1707.01629* (2017).

[7] De Cheng et al. "Person re-identification by multi-channel parts-based cnn with improved triplet loss function". In: *Proceedings of the iEEE conference on computer vision and pattern recognition*. 2016, pp. 1335–1344.

[8] Xiao Chu et al. "Multi-context attention for human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1831–1840.

[9] Peng Cui, Shaowei Liu, and Wenwu Zhu. "General knowledge embedded image representation learning". In: *IEEE Transactions on Multimedia* 20.1 (2017), pp. 198–207.

[10] Changxing Ding et al. “Multi-task learning with coarse priors for robust part-aware person re-identification”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[11] David Eigen and Rob Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2650–2658.

[12] Keren Fu et al. “Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 3052–3062.

[13] Haiyun Guo et al. “Two-level attention network with multi-grain ranking loss for vehicle re-identification”. In: *IEEE Transactions on Image Processing* 28.9 (2019), pp. 4328–4338.

[14] Bing He et al. “Part-regularized near-duplicate vehicle re-identification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3997–4005.

[15] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[16] Lingxiao He et al. “Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7073–7082.

[17] Lingxiao He et al. “Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 8450–8459.

[18] Lingxiao He et al. “Recognizing partial biometric patterns”. In: *arXiv preprint arXiv:1810.07399* (2018).

[19] Alexander Hermans, Lucas Beyer, and Bastian Leibe. “In defense of the triplet loss for person re-identification”. In: *arXiv preprint arXiv:1703.07737* (2017).

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. “In defense of the triplet loss for person re-identification”. In: *arXiv preprint arXiv:1703.07737* (2017).

[21] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[22] Max Jaderberg et al. “Spatial transformer networks”. In: *arXiv preprint arXiv:1506.02025* (2015).

[23] Hanyang Jin, Shenqi Lai, and Xueming Qian. “Occlusion-sensitive Person Re-identification via Attribute-based Shift Attention”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2021).

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7482–7491.

[25] Alex Kendall, Matthew Grimes, and Roberto Cipolla. “Convolutional networks for real-time 6-DOF camera relocalization. CoRR abs/1505.07427 (2015)”. In: *arXiv preprint arxiv:1505.07427* (2015).

[26] Pirazh Khorramshahi et al. “The devil is in the details: Self-supervised attention for vehicle re-identification”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 369–386.

[27] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. “Bilinear attention networks”. In: *arXiv preprint arXiv:1805.07932* (2018).

[28] Iasonas Kokkinos. “Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6129–6138.

[29] Ratnesh Kuma et al. “Vehicle re-identification: an efficient baseline using triplet embedding”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–9.

[30] Dangwei Li et al. “Learning deep context-aware features over body and latent parts for person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 384–393.

[31] Wei Li, Xiatian Zhu, and Shaogang Gong. “Harmonious attention network for person re-identification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2285–2294.

[32] Yuqi Li et al. “Deep joint discriminative learning for vehicle re-identification and retrieval”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 395–399.

[33] Zechao Li and Jinhui Tang. "Weakly supervised deep metric learning for community-contributed image retrieval". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 1989–1999.

[34] Xianmin Lin et al. "Occlusion Based Discriminative Feature Mining for Vehicle Re-identification". In: *International Conference of Pioneering Computer Scientists, Engineers and Educators*. Springer. 2020, pp. 246–257.

[35] Hongye Liu et al. "Deep relative distance learning: Tell the difference between similar vehicles". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2167–2175.

[36] Wei Liu et al. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[37] Xiaobin Liu et al. "Ram: a region-aware deep model for vehicle re-identification". In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2018, pp. 1–6.

[38] Xihui Liu et al. "Hydraplus-net: Attentive deep features for pedestrian analysis". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 350–359.

[39] Xinchen Liu et al. "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance". In: *IEEE Transactions on Multimedia* 20.3 (2017), pp. 645–658.

[40] Yihang Lou et al. "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3235–3243.

[41] Hao Luo et al. "Bag of tricks and a strong baseline for deep person re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.

[42] Hao Luo et al. "Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification". In: *IEEE Transactions on Multimedia* 22.11 (2020), pp. 2905–2913.

[43] Dechao Meng et al. "Parsing-based view-aware embedding network for vehicle re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7103–7112.

[44] Ishan Misra et al. “Cross-stitch networks for multi-task learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3994–4003.

[45] Hyun Oh Song et al. “Deep metric learning via lifted structured feature embedding”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4004–4012.

[46] Maxime Oquab et al. “Learning and transferring mid-level image representations using convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1717–1724.

[47] Jingjing Qian et al. “Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification”. In: *Measurement Science and Technology* 31.9 (2020), p. 095401.

[48] Qi Qian et al. “Fine-grained visual categorization via multi-stage metric learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3716–3724.

[49] Xuelin Qian et al. “Multi-scale deep learning architectures for person re-identification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5399–5408.

[50] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[51] Pengyuan Ren and Jianmin Li. “Factorized distillation: Training holistic person re-identification model by distilling an ensemble of partial reid models”. In: *arXiv preprint arXiv:1811.08073* (2018).

[52] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “” GrabCut” interactive foreground extraction using iterated graph cuts”. In: *ACM transactions on graphics (TOG)* 23.3 (2004), pp. 309–314.

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[54] Pierre Sermanet et al. “Overfeat: Integrated recognition, localization and detection using convolutional networks”. In: *arXiv preprint arXiv:1312.6229* (2013).

[55] Yantao Shen et al. “Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1900–1909.

[56] Yifan Sun et al. “Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 393–402.

[57] Ziruo Sun et al. “CFVMNet: A Multi-branch Network for Vehicle Re-identification Based on Common Field of View”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 3523–3531.

[58] Marvin Teichmann et al. “Multinet: Real-time joint semantic reasoning for autonomous driving”. In: *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2018, pp. 1013–1020.

[59] Shangzhi Teng et al. “Multi-view spatial attention embedding for vehicle re-identification”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[60] Shangzhi Teng et al. “Scan: Spatial and channel attention network for vehicle re-identification”. In: *Pacific Rim conference on multimedia*. Springer. 2018, pp. 350–361.

[61] Jonas Uhrig et al. “Pixel-level encoding and depth layering for instance-level semantic labeling”. In: *German Conference on Pattern Recognition*. Springer. 2016, pp. 14–25.

[62] Guanshuo Wang et al. “Learning discriminative features with multiple granularities for person re-identification”. In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 274–282.

[63] Wenguan Wang et al. “Salient object detection with pyramid attention and salient edges”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1448–1457.

[64] Zhongdao Wang et al. “Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 379–387.

[65] Nattachai Watcharapinchai and Sitapa Rujikietgumjorn. “Approximate license plate string matching for vehicle re-identification”. In: *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2017, pp. 1–6.

[66] Xiu-Shen Wei et al. “Selective convolutional descriptor aggregation for fine-grained image retrieval”. In: *IEEE Transactions on Image Processing* 26.6 (2017), pp. 2868–2881.

[67] Kilian Q Weinberger and Lawrence K Saul. “Distance metric learning for large margin nearest neighbor classification.” In: *Journal of machine learning research* 10.2 (2009).

[68] Yandong Wen et al. “A discriminative feature learning approach for deep face recognition”. In: *European conference on computer vision*. Springer. 2016, pp. 499–515.

[69] Sanghyun Woo et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[70] Xun Yang, Meng Wang, and Dacheng Tao. “Person re-identification with metric learning using privileged information”. In: *IEEE Transactions on Image Processing* 27.2 (2017), pp. 791–805.

[71] Xun Yang et al. “Enhancing person re-identification in a self-trained subspace”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13.3 (2017), pp. 1–23.

[72] Xun Yang et al. “Enhancing person re-identification in a self-trained subspace”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13.3 (2017), pp. 1–23.

[73] Hantao Yao et al. “Deep representation learning with part loss for person re-identification”. In: *IEEE Transactions on Image Processing* 28.6 (2019), pp. 2860–2871.

[74] Dominik Zapletal and Adam Herout. “Vehicle re-identification for automatic video traffic surveillance”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 25–31.

[75] Peixi Zhang et al. “MGD: mask guided de-occlusion framework for occluded person re-identification”. In: *International Conference on Intelligent Science and Big Data Engineering*. Springer. 2019, pp. 411–423.

[76] Xiaofan Zhang et al. “Embedding label structures for fine-grained feature representation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1114–1123.

[77] Xinyu Zhang et al. “Part-guided attention learning for vehicle instance retrieval”. In: *IEEE Transactions on Intelligent Transportation Systems* (2020).

[78] Zhizheng Zhang et al. “Densely semantically aligned person re-identification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 667–676.

[79] Yanzhu Zhao et al. “Structural analysis of attributes for vehicle re-identification and retrieval”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.2 (2019), pp. 723–734.

[80] Heliang Zheng et al. “Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5012–5021.

[81] Qi Zheng et al. “Car re-identification from large scale images using semantic attributes”. In: *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2015, pp. 1–5.

[82] Wei-Shi Zheng et al. “Partial person re-identification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4678–4686.

[83] Zhedong Zheng, Liang Zheng, and Yi Yang. “Pedestrian alignment network for large-scale person re-identification”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.10 (2018), pp. 3037–3045.

[84] Zhun Zhong et al. “Re-ranking person re-identification with k-reciprocal encoding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1318–1327.

[85] Feng Zhou and Yuanqing Lin. “Fine-grained image classification by exploring bipartite-graph labels”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1124–1133.

[86] Yi Zhou and Ling Shao. "Vehicle re-identification by adversarial bi-directional lstm network". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 653–662.

[87] Yi Zhou and Ling Shao. "Viewpoint-Aware Attentive Multi-View Inference for Vehicle Re-Identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6489–6498.

[88] Jiaxuan Zhuo et al. "Occluded person re-identification". In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2018, pp. 1–6.