Seeing Crucial Parts: Vehicle Model Verification via A Discriminative Representation Model

LIQIAN LIANG, School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
CONGYAN LANG, Corresponding author, School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
ZUN LI, Beijing Jiaotong University, China
JIAN ZHAO, Institute of North Electronic Equipment, China
TAO WANG, Beijing Jiaotong University, China
SONGHE FENG, Beijing Jiaotong University, China

Widely-used surveillance cameras have promoted large amounts of street scene data, which contains one important but long-neglected object: vehicle. Here we focus on the challenging problem of vehicle model verification. Most previous works usually employ global features (*e.g.*, fully-connected features) to further perform vehicle-level deep metric learning (*e.g.*, triplet-based network). However, we argue that it is noteworthy to investigate the distinctiveness of local features and consider vehicle-part-level metric learning by reducing the intra-class variance as much as possible. In this paper, we introduce a simple yet powerful deep model, *i.e.*, enforced intra-class alignment network (EIA-Net), which can learn a more discriminative image representation by localizing key vehicle parts and jointly incorporating two distance metrics: vehicle-level embedding and vehicle-part-sensitive embedding. For learning features, we propose an effective feature extraction module which is composed of two components: Regional Proposal Network (RPN)-based network and Part-based CNN. RPN-based network is used to define key vehicle regions and aggregate local features on these regions, while Part-based CNN offers supplementary global features for RPN-based network. The fusion features learned by feature extraction module are cast into deep metric learning module. Especially, we derived an enforced intra-class alignment loss (EIAL) by re-utilizing key vehicle part information to enhance reducing intra-class variance. Furthermore, we modify the coupled cluster loss (CCL) to model the vehicle-level embedding by enlarging the inter-class variance while shortening intra-class variance. Extensive experiments over benchmark datasets VehicleID and CompCars have shown that the proposed EIA-Net significantly outperforms the state-of-the-art approaches for vehicle model verification. Furthermore, we also conduct comprehensive experiments on vehicle Re-ID datasets, *i.e.*, VehicleID and VeRi776, to validate the generalization ability effectiveness of our proposed method.

CCS Concepts: • **Computing methodologies → Computer vision**.

Authors' addresses: Liqian Liang, School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China, 100044, lqliang@bjtu.edu.cn; Congyan Lang, Corresponding author, School of Computer and Information Technology, Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China, 100044, cylang@bjtu.edu.cn; Zun Li, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China, 100044, 16112072@bjtu.edu.cn; Jian Zhao, Institute of North Electronic Equipment, 226 North Fourth Ring Road, Haidian District, Beijing, China, 100191, zhaojian90@u.nus.edu; Tao Wang, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China, 100044, twang@bjtu.edu.cn; Songhe Feng, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China, 100044, shfeng@bjtu.edu.cn.

0004-5411/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

Additional Key Words and Phrases: Vehicle model verification, vehicle re-identification, image representation learning, deep metric learning

ACM Reference Format:
Liqian Liang, Congyan Lang, Zun Li, Jian Zhao, Tao Wang, and Songhe Feng. 2018. Seeing Crucial Parts: Vehicle Model Verification via A Discriminative Representation Model. *J. ACM* 37, 4, Article 111 (August 2018), 22 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recently, the widespread surveillance cameras have promoted large amounts of street scene data, which contains one important object: vehicle, and therefore posed vehicle-related tasks such as vehicle image retrieval, vehicle **re-id**entification (Re-ID) and vehicle model verification. Actually, all these three tasks can be seen as a sub-problem of image retrieval. In this paper, we highlight the problem of vehicle model verification, aiming to identify whether two vehicles belong to the same model. Note that the categories of vehicle models are normally organized in three-level hierarchy architecture, *i.e.*, make, model and released year [58]. Hence it is more challenging to address this problem in view of the subtle differences among vehicles with different released years belonging to the same model and make. Meanwhile, vehicle model verification is critical for public security and intelligent transportation system, *e.g.*, when the vehicle plates are occluded due to the viewpoint change or different location of the surveillance cameras, the track for certain vehicle model becomes more important in transportation regulation or accident investigation.

Most previous vehicle-related works target at solving the problem of vehicle Re-ID [25][2]. As far as we know, there is nearly no literature about vehicle model verification [25][58] in computer vision community. Unlike vehicle Re-ID that identifies the same vehicle across different camera views, vehicle model verification concentrates on vehicle-model-level identification instead of instance-level identification. Though vehicle Re-ID provides finer recognition, it is hungry for more fine-grained details to confirm the vehicle identity such as customized painting or decorations. However, over-attention on subtleties is inadequate under some circumstances where such tiny differences can be easily ignored, *e.g.*, in video tracking with low resolution and motion blur.

Advances towards vehicle-related tasks either focus on discriminative feature learning or deep metric learning to resolve the misalignment issue due to large variation of viewpoint and illumination, heavy occlusion, background noise, *etc*. For discriminative feature learning, early works usually adopt the fully-connected features of CNN as global representations. We argue that such global features only emphasize semantic information such as color but lack the capacity to handle the spatial relationships and misalignment scenarios caused by large inter-class variance and small intra-class variance as illustrated in Fig. 1. There also emerge several studies that explore local features to alleviate the aforementioned issue by introducing extra data with annotations, such as keypoints [53], viewpoints [67], bounding boxes [14]and segmentation masks [28, 34]. However, those methods which utilize keypoints or viewpoints ignore the fact that the discriminative cues usually appear in specific regions of vehicles and suffer from the challenge of subtle inter-instance discrepancy of similar vehicles. The methods with segmentation annotation rely on more explicit labelling and the existing bounding box based methods normally utilize multiple branches to perform local feature extraction for each detected region, which is less flexible.

Deep metric learning has proved its potentials in various tasks such as face recognition [38], person/ vehicle Re-ID [2, 5, 25, 57, 59]. In general, the purpose of deep metric learning is to pull the images with same labels closer meanwhile push the images with different labels far away, like classic triplet loss [38, 54], **c**oupled **c**luster **l**oss (CCL)[25], *etc.* However, these deep metric learning methods only rely on object-level embedding (*e.g.*, features extracted from different vehicle images), which overlook the impact of object-part-level embedding (*e.g.*, features extracted from different

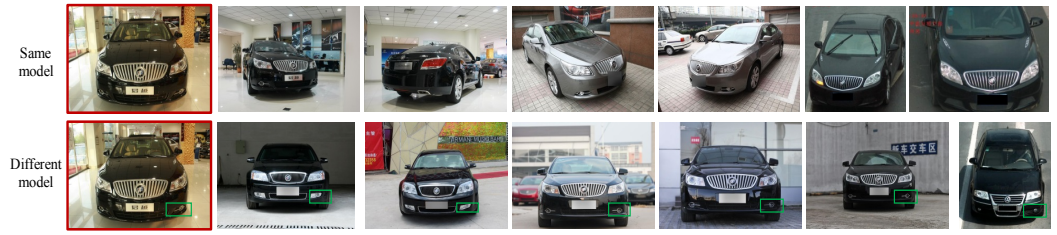

Fig. 1. The illustration of intra-class and inter-class variance.Some example samples are shown here to represent the complex scenarios for the problem, *i.e.*, the same vehicle models can have large appearance difference due to the variations of illumination or viewpoint, while the appearance of different vehicle models are sometimes nearly identical. However, we can identify the same vehicle model by its local part characteristics, such as the headlights denoted by green bounding boxes.

key vehicle parts). In other words, most deep metric learning methods employ the final image representations to model the difference between inter-class variance and intra-class variance, and few of them take advantage of local features from object-part-level to reduce intra-class variance as far as possible, which may be beneficial for capturing more details within the same class.

Motivated by the above facts, we intend to derive a more flexible feature extractor that can jointly learn global features and local features, since global features focus more on capturing semantic information while local features values the importance of vehicle parts that may possess more distinguishable clues. Towards these two features, we further propose an enforced intra-class alignment loss function and a modified CCL function that can embed vehicle-part-level and vehicle-level features to latent spaces, respectively. To this end, we propose a simple yet powerful deep model, *i.e.*, enforced intra-class alignment network (EIA-Net) which is composed of two crucial modules: feature extraction module based on key vehicle part localization and deep metric learning module.

Inspired by the observations that different vehicle models can be distinguished via key vehicle parts (*e.g.*, the headlights as shown in Fig. 1), we assume that the aggregation of local features from precisely-detected key vehicle parts possess more discriminative abilities to describe the differences of vehicle appearances. Thus for feature extraction, we first localize crucial vehicle regions to further construct a powerful image descriptor as the input of later deep metric learning module. In order to attain a compact and fixed-length image representation regarding key vehicle parts, we propose an effective way to perform local feature aggregation inspired by the construction way of R-MAC descriptor [12]. Considering that local features cannot model the object-level context information, we further propose to learn complementary global features in the form of Part-based CNN [6]. Finally, our feature extraction module learns more distinguishable local features and global features simultaneously to form the final feature construction. Note that we utilize an extra database [1] where the images are selected from VOC2012 Dataset and **r**egional **p**roposal **n**etwork (RPN) [37] for key vehicle parts detection.

In this paper, we also consider to jointly incorporate vehicle-level embedding and **v**ehicle-**p**art-**s**ensitive **e**mbedding (VPSE). For vehicle-level embedding, we model different-hierarchy inter-class variance by introducing different margins to modify the original CCL, since the vehicle models are more difficult to verify when them belong to the same model but different released years. To further shorten intra-class variance, we propose an **e**nforced **i**ntra-class **a**lignment **l**oss (EIAL) to

[1]Researchers interested in this vehicle part detection database may consult the author for approval of Australian Centre for Robotic Vision, School of Computer Science in the University of Adelaide.

perform VPSE by reutilizing the local features of detected key vehicle parts. It is noteworthy that our proposed VPSE is essentially a more fine-grained distance metric which maps the representations of the same key vehicle part of image pairs into the same feature space, aiming to pull them closer. Nevertheless, VPSE may cause overfitting to deep network, hence we consider a simple strategy to relax the constraints which will be presented in detail in Section 3. Extensive experiments over benchmarks, *i.e.*, CompCars [58] and VehicleID [25], demonstrate that our proposed EIA-Net significantly outperforms the state-of-the-art methods. Furthermore, we also conduct comprehensive experiments on vehicle Re-ID datasets, *i.e.*, VehicleID [25] and VeRi-776 [27], to validate the generalization ability effectiveness of our proposed method.

The main contributions of this paper are summarized as three-fold.

1. We propose a simple yet powerful deep model EIA-Net to address the problem of vehicle model verification. EIA-Net learns a more discriminative image representation fused by local features based on detected key vehicle parts and global features extracted by Part-based CNN.

2. We propose an effective deep metric leaning method which incorporates two distance metrics: vehicle-level embedding and vehicle-part-sensitive embedding (VPSE).

3. Comprehensive evaluations on benchmark datasets verify the superiority of EIA-Net over the state-of-the-arts and the robustness of EIA-Net for vehicle Re-ID.

The rest of the paper is organized as follows. Related works are discussed in Section 2. In Section 3, we propose a unified deep network for vehicle model verification that comprises two crucial modules: feature extraction module and deep metric learning module. Experimental results on mainstream datasets are described in detail in Section 4 and Section 5 conclude the main idea of our work.

2 RELATED WORK

In this section, we discuss the previous studies in relation to our method, which mainly focus on three aspects: vehicle-related works, deep metric learning and other related works such as aggregation of local features used in image retrieval.

2.1 Vehicle-related Works.

Advanced by deep learning networks, recent vehicle-related works have been greatly facilitated, *e.g.*, vehicle model classification [17, 20, 24, 58, 62], vehicle attribute prediction [17, 55], vehicle image retrieval [11], and vehicle Re-ID [2, 8, 25, 26]. For early vehicle model classification, Hsiao *et al.* [17] Lin *et al.* [24] and Krause *et al.* [20] focused on the image alignment, fitting and constructing image representations of 3D vehicle model respectively. All these works are restricted to a small number of vehicle models until Yang *et al.* [58] built a large-scale dataset CompCars which contains 431 vehicle models. On top of CompCars, model classification experiments are performed on two settings: using the entire car images or the car parts, both of which adopt the logistic function to fine-tune the Overfeat model [40].

Many vehicle-related efforts are devoted to vehicle image retrieval or vehicle Re-ID lately. Technically, there are two core concerns about these tasks: feature extraction for a more discriminative descriptor and deep metric learning for embedding across images. For vehicle image retrieval, in [11], vehicle attributes and colors are identified first and then used to further image retrieval based on these recognized vehicle attributes. Here we briefly review the vehicle Re-ID methods as follows. Most existing methods can be roughly categorized into four groups according to the extraction ways of features: (1) **Hand-crafted features.** Before the popularity of deep learning, most methods emphasize this kind of features to filter the noise and improve the discriminativeness, like LOMO [23] and BOW-CN [63]. (2) **Global features extracted by CNN.** Early deep learning based methods usually rely on fully connected layers of CNN to learn global visual features, including GoogLeNet [58], Siamese-CNN [41], NuFACT [27], RAM [29], MLSL [1], and FDA-Net [32]. (3) **Multi-view**

features. This type of features are normally extracted from different sources such as license plate, spatio-temporal context, *etc*, of which the outstanding ones include OIFE+ST [53], Siamese+ST [41], and PROVID [27]. (4) **Fusion of global and local features.** There emerge several extra information (*e.g.*, keypoints, viewpoints, bounding boxes of parts and segmentation masks) based methods like ours. The representative approaches are numerated as follow: OIFE [53], VAMI [67], C2F [13], EALN [31], AAVER [19], PRN [14], PAMTRI [50], PVEN [34], PCRNet [28], and VehicleNet [64]. There also emerge several methods that focus on the deep metric learning. Liu *et al.* [25] addressed the re-ID task by posing a deep relative distance learning (DRDL) method that employs coupled clusters loss (CCL) function. Bai *et al.* [2] propose a deep metric learning method, *i.e.*, group-sensitive-triplet embedding (GS-TRE) in which intra-class variance is modeled by incorporating an intermediate representation "group" between vehicle samples under a triplet-based network setting.

There is nearly no literature about our concerned vehicle model verification [25, 58] compared with the popular face verification [38, 46–49]. Yang *et al.* [58] followed the pipeline of state-of-the-art face verification method [47], where the features are extracted by CNN and then combined with the well-known Joint Bayesian method [3]developed for face verification. Liu *et al.* [25] performed the vehicle verification task by using the DRDL model, which is inspired by deep metric learning [7, 18, 42, 45] and Siamese network [12, 36, 38, 46].

We found that of all the approaches presented above, nearly all of them overlooks to take advantage of vehicle-part-level features. They naturally extract the fully-connected layer features, *i.e.*, global descriptors, to recognize vehicle attributes or learn a distance metric embedding. In CompCars, they proved that using the "taillight" part can reach the best accuracy over all the other car parts. However, their car parts need to be segmented manually and the back view of vehicles only occupy a small percentage of the whole views. Therefore, we utilize state-of-the-arts object detection methods to automatically locate key vehicle parts in the presence of various viewpoints, illumination *etc.*, for investigating a more discriminative image representation aggregated by local features based on these vehicle parts. In general, the framework for vehicle model verification is similar to that for vehicle Re-ID or retrieval. Prior works focus on either capturing a powerful feature or deep metric learning, few works [2, 25] put emphasis on jointly considering these two tasks in a unified framework. However, their works still ignore the informative local features of key vehicle parts. In the following subsection, we will discuss the detailed impact of different deep metric learning methods.

2.2 Deep Metric Learning

The aim of deep metric learning method is to maximize inter-class distances meanwhile minimize the intra-class distances. So far, the pioneering triplet loss function [54] which is originally used for the problem of nearest neighbor classification, has proved its superiority and later widely used in face recognition [38], fine-grained object recognition [52], pedestrian Re-ID [5, 57, 59] and vehicle Re-ID [2, 25]. On the basis of triplet loss, more works are devoted to explore deeper and richer inter-class relationships [9, 59, 61, 65] or develop triplet-based variants[5, 22, 25]. Yang *et al.* [59] leveraged privileged information and large amounts of unlabeled samples as complements for constructing distance metric. In [61], Zhang *et al.* proposed to take hierarchical inter-class relationship via the injection of multiple labels, *i.e.*, different models, brands, manufactured years, as prior knowledge and learn a feature embedding based on this hierarchical inter-class relationship. Lin *et al.* [65] utilized bipartite-graph labels to model rich inter-class relationships based on multiple sub-category components. In [9], a general knowledge graph was proposed to capture the relations of class labels, then a regularized regression model is employed to jointly optimize the image representation learning and graph embedding. Chen *et al.* [5] proposed a quadruplet network to improve the generalization capability of original triplet network. In [22], Li *et al.* explored how to learn a distance metric

to incorporate corresponding semantic information via user-provided tags and further boosted the performance of tag-based image retrieval. In [25], coupled cluster loss was proposed to improve the triplet loss by introducing a cluster center point of positive samples thus avoiding randomly selecting anchor samples. Bai *et al.* [2] modeled intra-class variance and inter-class variance simultaneously. Specifically, they learned the distance between an anchor sample and other group anchors with the same Vehicle ID to enforce the preservation of the intrinsic attributes of the instances.

Most previous works only focus on optimizing the inter-class distance or the difference between the inter-class variance and intra-class variance, yet ignore learning more discriminative local features within the same class, which is especially useful for dealing with large intra-class variance. Though Bai *et al.* imposed a more fine-grained local structure constraints within a class into deep metric learning to model intra-class variance, it remains to adopt vehicle-level image representations. In our paper, to shorten the vehicle-part-level intra-class variance, we directly locate key vehicle parts to investigate the subtleties of key vehicle parts and utilize the local features to enforce class alignment for each key vehicle part.

2.3 Other Related Works

Aggregation of local convolutional descriptors. Over the past few years, CNN fully-connected layer feature has shown its superiorities in image classification [21], recognition [43] and semantic segmentation [4, 30]. However, in the field of image instance retrieval, many works prove that the activations of convolutional layers outperform fully-connected layer features, thus resulting in the improvement of aggregation methods on convolutional features, such as spatial max-pooling (MAC) [10] and sum-pooling (SPoC) [16]. On the basis of MAC, Tolias *et al.* [51] introduced R-MAC, an image representation by aggregating the convolutional features of a fixed layout square regions. Gordo *et al.* [12] put forward to exploit RPN [37] to generate a set of proposals to replace the fixed regions, considering the rigid grid of regions can easily deviate from the interesting area. Inspired by this, our model also adopts the RPN to locate the key vehicle parts to get the local convolutional responses.

Part-based CNN. Part-based CNN is commonly used for reinforcing distinctiveness of image representations on account of considering both the features of full body and the features of the body parts [6, 60, 69]. We adopt part-based CNN mainly to make up for the lacking consideration of global features. If we only take account into the local features of key part vehicle regions, vehicle model verification can become more difficult for two vehicles with large appearance differences.

3 PROPOSED METHOD

This section introduces our proposed method, **e**nforced **i**ntra-class alignment network (EIA-Net), for vehicle model verification. First, we present our overall framework that is composed of two crucial modules: feature extraction module and deep metric learning module. Specially, feature extraction module utilizes state-of-the-art detection method to locate key vehicle parts, then learn an image descriptor which incorporates local features based on these parts and global features generated by Part-based CNN. We further propose a deep metric learning module that incorporates vehicle-part-sensitive embedding and vehicle-level variance embedding via modeling intra-class alignment for each vehicle part and inter-class variance, respectively. To this end, we posed corresponding **e**nforced **i**ntra-class **a**lignment **l**oss (EIAL) and modified **c**oupled **c**luster **l**oss (CCL) for a more powerful distance metric.

3.1 The Overall Framework

As illustrated in Fig. 2, our overall EIA-Net takes positive set (images with the same vehicle ID) and negative set (images with the different vehicle ID from those in positive set) as input, then flow

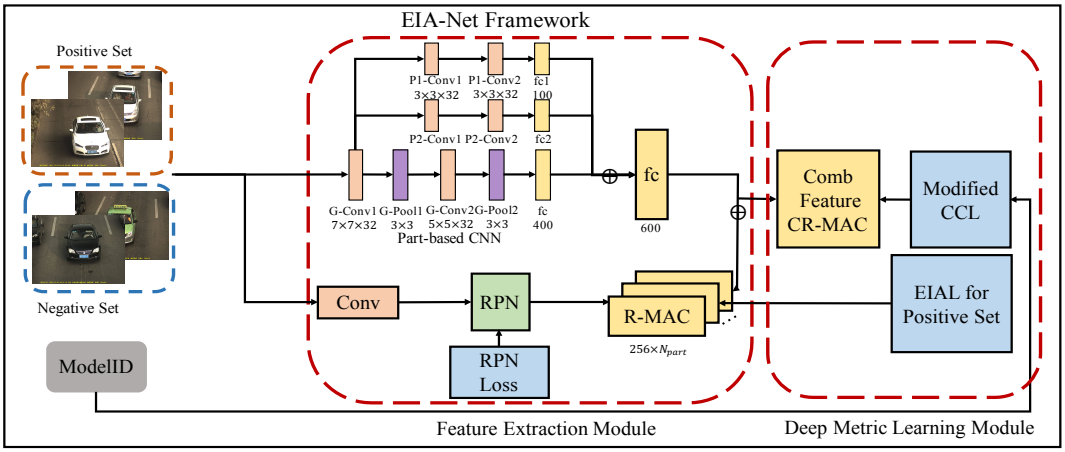

Fig. 2. This figure depicts the overall framework and workflow of EIA-Net. The entire network consists of two modules: feature extraction module and deep metric learning module. The final image representation are concatenated by R-MAC descriptors and the features extracted by Part-based CNN. Then the network is trained by EIAL and modified CCL loss function, which is used for learning distance metric among images.

through two sub-nets for later feature extraction, finally outputs the embedding jointly fine-tuned by modified CCL and EIAL. For feature extraction, we provide the fusion of two features: global features and aggregation of local features based on key vehicle parts extracted by Part-based CNN and **r**egional **p**roposal **n**etwork (RPN)-based network respectively. The Part-based CNN includes one global convolutional layer as well as the following global branch and Part-based branches for generating one global representation and part-based representations. Here we adopt one global branch and two part-based branches. Specifically, the feature maps of the first convolutional layer in Part-based branches are equal parts divided by feature maps of the global convolutional layer. The kernel size and the arrangement of different layers are depicted in Fig. 2. The features extracted by Part-based CNN are treated as complementary global features and coarse local features for the aggregation of local descriptors based on key vehicle parts, which is implemented by RPN-based network. We employ the similar architecture of heads and RPN in [37], then fine-tuned RPN on our extra dataset to generate proposals of key vehicle parts. We construct the local features of key vehicle parts by modifying the formulation of R-MAC. The final features **c**ombined **R-MAC** (CR-MAC) are the concatenation of the features extracted by Part-based CNN and RPN-based Network. On the basis of powerful image representations, we perform deep metric learning which incorporates vehicle-level variance and vehicle-part-sensitive variance simultaneously. In particular, modified CCL focus on modeling the vehicle-level variance by computing the difference of center point/positive pair and center point/negative pair (the definition of center point is described in Subsection 3.3). That is to enlarge the difference between the inter-class variance and intra-class variance. Meanwhile, EIAL mainly aims to reduce the intra-class variance as much as possible by introducing the vehicle-part-level feature alignment, which formulates the difference among the local features belonging to the same key vehicle part between two positive images. Feature extraction module will be described in Subsection 3.2 while vehicle-level and vehicle-part-sensitive deep metric learning will be presented in Subsection 3.3 and Subsection 3.4.

3.2 Learning to Extract Features of Key Vehicle Parts

RPN revisited. RPN [37], the core component of Faster-RCNN, is a fully convolutional network for generating region proposals with high confidence, in which both the coordinates of object bounding box and objectiveness scores are predicted at each location. Actually, RPN plays the role as "attention" mechanism to tell the later object detection network where to look, which is implemented by "anchors", *i.e.*, a set of rectangular boxes associated with different scales and aspect ratios. Then the features are fed into two sibling fully-connected layers connected with one multi-task loss function to predict regression coordinates and box-classification scores. Therefore RPN possess the capacity to locate the interested regions. Inspired by this property of RPN, we utilize it in our framework as the detection backbone for localizing key vehicle parts (such as car light, back mirror, logo *etc.*). Based on these localized proposals, we first perform non-maximum suppression to locate each vehicle part to later local feature extraction. Here we adopted the default value, 3 scales and 3 aspect ratios, thus yielding 9 anchors in total for each location.

R-MAC descriptor revisited. How to effectively construct local features on localized vehicle parts is another crucial consideration in feature extraction module. Motivated by deep-based content instance image retrieval researches, we observe that activations of the last convolutional layer [16][10] achieves remarkable performance compared to fully-connected features, since activations reflects the most distinguishable regions in the overall images. Of all these methods, **r**egional **m**aximum **a**ctivation of **c**onvolutions (R-MAC) [51] proves to be competitive and more easily implemented. In order to construct R-MAC descriptors, we need to start from constructing **m**aximum **a**ctivation of **c**onvolutions (MAC). Given an input image, let the activations of an intermediate convolutional layer be denoted as a 3D tensor of $W \times H \times K$ dimensions, where W, H and K represent the width, height and the number of channels of output feature maps. The 3D tensor can also be represented as a set of 2D feature channel responses $\Phi = \{\Phi_i\}, i = 1 \dots K$. Then the spatial max-pooling is operated over all the locations of each 2D feature responses Φ_i to form the MAC as below.

$$f_\Omega = \{f_\Omega^{(i)} | i = 1, 2 \dots K\}, with f_\Omega^{(i)} = \max_{p \in \Omega} \Phi_i(p), \tag{1}$$

where Ω represents the pixel set of the feature maps and $\Phi_i(p)$ is the response at a particular position. Now assume that an image is composed of R regions with different scales, the max-pooling is operated on these regions instead of the whole feature map. The R-MAC descriptor can be formulated as below.

$$f'_\Omega = \{f_{\Omega_j}^{(i)} | i = 1, 2 \dots K, j = 1, 2, \dots N_{part}\}, with f_{\Omega_j}^{(i)} = \max_{p \in \Omega_j} \Phi_i(p), \tag{2}$$

where Ω_j represents the pixel set of j_{th} region within the convolutional feature maps and N_{part} indicates the number of key vehicle parts. In original construction way of R-MAC descriptor, the regions are normally fixed and the dimension of R-MAC is equal to K through the sum collection process of regional values in the same feature map. In our paper, to enhance the distinguishable ability of local features and further perform the enforced intra-class alignment, we automatically detect N_{part} key vehicle regions then concatenate $f_{\Omega_j}^{(i)}$ along the channel axis to form the modified R-MAC, *i.e.*, f_{Ω_j} for each vehicle part. If one vehicle part is not detected, we employs zero vector as padding. When involving loss function computation, we only calculate the differences among co-detected key vehicle parts. Notice that the modified R-MAC is performed on the last convolutional layer.

Construction of final image representations. The modified R-MAC based on detected key vehicle parts mainly focus on the local recognizing ability. Considering this, we also take context

information into account via Part-based CNN. The design of Part-based CNN basically follows the work of [6], but our Part-based CNN includes one full-body channel and two equal body-part channels instead of their four body-part channels. In other words, Part-based CNN can simultaneously extract the features from two-level: global features and coarse local features, which are learned from full images and two equally partitioned parts of images. As illustrated in Fig. 2, the full-body channel contains two convolutional layers and two pooling layers while each body-part channel contains only two convolutional layers. Then features are concatenated by features of the three channel-wise fully-connected layers. Finally, we concatenate the modified R-MAC descriptors by the order of vehicle parts and features extracted in Part-based CNN to form the final image representations, the dimension of which is the sum of modified R-MAC $K \times N_{part}$ and number of channels of fully-connected layer C in Part-based CNN. Here we adopt the architecture of RPN and original Part-based CNN, thus K equals to 256 and C equals to 600. The **c**ombined R-MAC is herein called CR-MAC in our paper, later cast into our overall network trained jointly by modified CCL and EIAL.

3.3 Modeling Vehicle-level Variance

Coupled cluster loss revisited. The concept of CCL is extended from triplet loss, where the inputs are triplet units $\{\langle x^a, x^p, x^n \rangle\}$ which indicate an anchor, a positive image that belongs to the same identity as anchor does and a negative image respectively. The triplet units should satisfy the constraint that the distance between the anchor and the negative image exceeds at least the defined margin than the distance between the anchor and the positive image. While in CCL, although the main idea behind it is similar to triplet loss, there are only positive and negative sets rather than triplet units with extra anchors. CCL replaces the anchor with the center point, *i.e.*, the mean value of all positive samples, to compare the relative distance relationship, which can be reflected as

$$\| f(x_i^p) - c^p \|_2^2 + \alpha \leq \| f(x_j^n) - c^p \|_2^2 \ \forall 1 \leq i \leq N^p and 1 \leq j \leq N^n, \tag{3}$$

where c^p denotes the center point, α is a scalar that represents the minimum margin between matched and mismatched pairs and $f(*)$ represents the transformation from the input images to extracted features. For each batch, CCL is denoted as

$$L(W, X^p, X^n) = \sum_i^{N_p} \frac{1}{2} \max\{0, \| f(x_i^p) - c^p \|_2^2 + \alpha - \| f(x_*^n) - c^p \|_2^2\}, \tag{4}$$

where x_*^n is the nearest negative samples to the center point and N_p indicates the number of positive sets in our paper. According to the loss function, the partial derivatives of the positive samples and the nearest negative sample can be computed easily so that the convergence of the whole network can be guaranteed. Compared to triplet loss, CCL is far more reliable unlike the randomly-selected anchors.

Modified coupled clusters loss. CCL is designed specially for the problem of vehicle Re-ID. To suit for the problem of vehicle model verification, we modify the CCL function since there is no need for justifying whether two vehicles are of the same identity. In the original CCL, the positive set of the input indicates the images that are labeled with the same identity, *i.e.*, they belong to the same vehicle. Here we replace the positive set with the images of the same vehicle model of certain specific released year. Therefore, the center point denotes the mean value of all the positive samples of the same released year from the same model. In addition, the vehicle models of different released years are similar to each other since the manufacture usually change few vehicle parts, thus more challenging to be verified in contrast to the vehicles of different models even different makes. So we regard the image set of the same vehicle model but of different released years as hard training

examples. Such hard training samples contribute more to the back propagation process than the easy ones because the partial derivative can vary at a greater value during one iteration. To this end, we modify the CCL with only one scalar that manifests the margin between the relative distance by utilizing multiple margins to apply for the hard and easy training samples. Especially we adopt the larger margin for the negative images that are of different released year but the same model. The modified CCL function is defined as

$$L(W, X^p, X^n) = \sum_{i}^{N_p} \frac{1}{2} \max\{0, \| f(x_i^p) - c^p \|_2^2 + \alpha - g(x_*^n)\}, \tag{5}$$

$$g(x_*^n) = \begin{cases} \| f(x_*^n) - c^p \|_2^2, \text{when } x_*^n \text{ is a negative sample of different models,} \\ \| f(x_*^n) - c^p \|_2^2 - \beta, \text{when } x_*^n \text{ is a negative sample of the same model,} \end{cases} \tag{6}$$

where $g(*)$ function takes the different value as input dependent on whether x_*^n is a negative sample of different models and $\alpha + \beta$ is the value of the larger margin.

The differences between the original CCL and modified CCL are summarized as below:

1. The positive set represents the images which belong to the same vehicle models of the same released year rather than the image set where the images belong to the same vehicle;

2. The margin in the original CCL function is replaced by multiple margins, so that it can better represent the divergence of the loss when dealing with hard or simple examples.

3.4 Modeling Enforced Vehicle-part-sensitive Variance

As mentioned in the former subsection, CCL function mainly focus on vehicle-level distance metric learning via the aggregation of local features and global features, extracted by RPN-based network and Part-based CNN respectively. However, CCL loss overlook the may remain insensitive to subtleties among the samples of different models, and it only models the difference between interclass variance and intra-class variance. To further pull the positive samples closer and enhance the discriminative ability of deep metric learning, we consider vehicle-part-sensitive embedding by putting a stronger constraint on key vehicle part alignment. That is based on the observation that key vehicle part contains more important information targeting at more fine-grained recognition or verification. Inspired by domain transfer or adaption [33][56] where **s**upervised **d**omain **a**daptation (SDA) method can ensure features of the same class from different domains to be mapped nearby, we model key vehicle part label information for our vehicle verification by class alignment.

In SDA, class alignment is only performed on object-level class labels between two domains. Here we adapt this alignment to vehicle model verification at the vehicle-part-level. Specifically, for every pair in positive sets, let the modified R-MAC feature vectors for all the key vehicle part proposals be $X_{R_i}^{(j)}$, where R_i denotes the R-MAC feature vector of the i_{th} key vehicle part and j represents the j_{th} sample in the positive set. The **e**nforced **i**ntra-class **a**lignment **l**oss (EIAL) is as follows.

$$\mathcal{L}_A = \sum_{j=1}^{N_{batch}} \sum_{i=1}^{N_{part}} \phi(X_{R_i}^{(j)}, X_{R_i}^{(k)}), j \neq k, \tag{7}$$

where ϕ, N_{batch} and N_{part} indicate the distance measure function, batchsize and the number of key vehicle parts respectively. We adopt the squard Euclidean distance in our experiments. Moreover, it can be beneficial to constrain the number of parameters N_{part} and pairs used in VPSE to address overfitting problem. To relax the assumption that representations of the same key vehicle part tend to be more similar, we may sample one pair in one batch and the most discriminative vehicle parts to perform such alignment. The detailed analysis is presented in Section IV.

4 EXPERIMENTS

In this section, we first introduce three vehicle-related datasets, *i.e.*, CompCars [58], VehicleID [25], VeRi776 [27] and VERI-Wild [32], and evaluation metrics for vehicle model verification as well as vehicle Re-ID. Then we elaborate the network settings and baseline design in Subsec. 4.2. In Subsec. 4.3, we provide detailed analysis on the influence of crucial components of our EIA-Net, including feature extraction module, EIAL, and modified CCL. Concretely, feature extraction module is composed of RPN, R-MAC and Part-based CNN. Finally, we compare the vehicle model verification results of our EIA-Net and state-of-the-arts on CompCars and VehicleID in Subsec. 4.4 and vehicle Re-ID results on VehicleID and VeRi776 in Subsec. 4.5.

4.1 Vehicle-Related Datasets and Evaluation Metrics

Note that CompCars and VehicleID described below can serve for the problem of vehicle model verification and VehicleID and VeRi776 are for the vehicle Re-ID task. We also introduce our extra dataset that are designed for producing key vehicle part proposals.

CompCars dataset. The CompCars dataset contains the images captured from two scenarios, the web-nature and the surveillance-nature. In particular, there are a total number of 136,727 images with entire cars and 27,618 images with only car parts in the web-nature, where most of them are annotated with five attributes (*maximum speed, displacement, number of doors, number of seats and type of car*) and five viewpoints (*front, rear, side, front-side and rear-side*). In the surveillance-nature, there are 50,000 images captured in the front view. Each image is labeled with bounding box and color of the car. A subset of CompCar Dataset which contains 78,126 images is partitioned into three parts, where Part-II and Part-III are utilized as the training set and test set for vehicle verification. The Part-II consists of 111 models with 4,454 images in total and Part-III comprises of 1,145 models with 22,236 images.

VehicleID dataset. The purpose of devising VehicleID Dataset is to solve the vehicle Re-ID task, thus the identity information of vehicles is specially collected. There are 26,267 vehicles existing in the total 221,763 images which are all captured in surveillance cameras from the front or back view. In addition, different from CompCars Dataset, only 250 vehicle models that most commonly appeared in real life are included in the Vehicle Dataset. Specifically, when involving the performance comparisons for vehicle model verification, we mainly focus on the images that have been annotated with model information. The number of images labeled with model information is 90,196. This dataset is split into two parts for training and testing, where the first part contains 47,558 images with vehicle model information and the second part consists of 42,638 images with vehicle model information.

VeRi-776 dataset. VeRi-776 is a classic vehicle ReID benchmark, containing 49,357 images of 776 vehicles which are collected from 20 cameras in the real traffic scenario under different viewpoints. The dataset is split into the training set that covers 576 vehicles and the test set that comprises the remaining 200 vehicles. VeRi-776 also provides several kinds of meta data such as the collected time and the location information.

VERI-Wild dataset. VERI-Wild contains 416,314 images identified by 40,671 vehicles captured from 174 cameras. The images are collected from various scenarios like complex backgrounds, varieties of viewpoints, different weather and illumination conditions. This dataset is randomly split into a training set with 277,797 images of 30,671 IDs and a testing set with 138,517 images of 10,000 IDs. The testing set is divided into three subsets according to data size, containing 3,000, 5,000, and 10,000 IDs, respectively.

Extra dataset. The training for the RPN requires the detailed coordinates and the labels of key vehicle parts, therefore we utilize our dataset selected from VOC2012 to achieve the target. We

annotate 17 vehicle parts that may contribute to the extraction of distinctive image features: *frontal, plate, wind glass, anu sign, car top window, back mirror, car light, logo, safe belt, paper box, light cover, hungs, entry license, carrier, newer sign, wheel and layon.* The images in our dataset are basically captured by surveillance cameras and the total number is 7,129. We split the dataset into the training set and test set which contains 4,940 and 2,189 images respectively. Unlike the CompCars and VehicleID Dataset, the images of this extra dataset are basically captured by surveillance cameras, which means there are a mass of difficult circumstances such as occlusion, motion blur and large change of viewpoints.

Evaluation metrics. For vehicle model verification, we adopt the widely-used average accuracy as the evaluation metric. As for vehicle Re-ID, we follow the prior works and employ two common evaluation metrics, *i.e.*, Rank-K and **m**ean **a**verage **p**recision (mAP). To be clearer, when given a ranking list, Rank-K indicates the probability that the positive image appears in the top K of this list, while mAP is the mean of the **a**verage **p**recision (AP) for all queries, of which each AP calculates the area under the recall-precision curve.

4.2 Implementation Details

Network settings. The training of our network starts from generating region proposals and extracting the local activations of the last convolutional layers. In the first part, we use the deep network (VGG16) of Simonyan *et al.* [44] pre-trained on the ImageNet ILSVRC challenges. Then we fine-tune the RPN using the RPN multi-task loss on our own vehicle dataset selected from VOC2012. During the second stage of the training, we fix the weights of RPN and use CCL to further fine-tune the rest of the network. The momentum of μ is set as 0.9 and weight decay λ equals to 2×10^{-4}. We start with a base learning rate at 0.01 and then drops by iteratively multiply 0.01 after every 10,000 batch iterations, where the batch size is set to 8. Loss weights of the upper branch, lower branch and the final one are 0.5, 0.5 and 1 respectively. All the setups of our network are implemented in deep learning framework pytorch.

Compared methods. Considering that the problem of vehicle model verification has not been sufficiently explored before, there exist only a few studies. For this task, we adopt two kinds of baselines: (1) **Combination of CNN fully-connected features and traditional classifier.** Here we follow the previous algorithms in [25] where they employ two traditional classifiers, *i.e.*, SVM and Joint Bayesian (named as "FC feature + SVM" and "FC feature + Joint Bayesian", respectively). Specifically, a deep convolutional network is first trained on Part-I data of CompCars as a feature extractor, then SVM or Joint Bayesian are applied on Part-II subset. (2) **Plain end-to-end deep network**. In order to validate the impacts of crucial components, we set a plain end-to-end deep network (named as "Plain DRDL") which is similar to DRDL method in [58], where only a two-branch network architecture and original coupled cluster loss are reserved.

For vehicle Re-ID, we compare our EIA-Net with existing methods which are roughly categorized into four groups according to the extraction ways of features: (1) **Hand-crafted features.** Before the popularity of deep learning, most methods emphasize this kind of features to filter the noise and improve the discriminativeness like LOMO [23]. (2) **Global features extracted by CNN.** Early deep learning based methods usually rely on fully connected layers of CNN to learn global visual features, including GoogLeNet [58], NuFACT [27], and FDA-Net [32]. (3) **Multi-view features.** This type of features are normally extracted from different sources such as license plate, spatio-temporal context, *etc*, of which the outstanding ones include OIFE+ST [53], and PROVID [27]. (4) **Fusion of global and local features.** There emerge several extra information (*e.g.*, keypoints, viewpoints, and bounding boxes of parts) based methods like ours. The representative approaches are numerated as follow: OIFE [53], VAMI [67], AAVER [19], PRN [14], and PAMTRI [50].

Fig. 3. The illustration of several image samples where the key vehicle parts are detected via the RPN. The first row indicates the original images in our dataset, and the second row indicates the detection results of key vehicle parts. From the detected images, we could see that the key vehicle parts cover the most of our interested regions, which are beneficial for extracting features with more discriminative capacities.

4.3 Ablation Study

In this subsection, we first prove the reliability of introducing RPN into our EIA-Net. Then we respectively validate the effectiveness of the core components in two crucial modules, *i.e.*, feature extraction module and deep metric learning module. Note that the vehicle model verification experiments in ablation study are conducted on CompCars and VehicleID and not distinguished to the released year level to make the comparison more credible, since baselines are not designed for recognizing the released year, either. Specifically, for VehicleID Dataset, we conduct the experiments on the subset with model label, which contains 47,558 and 42,638 images for training and testing.

Table 1. Accuracy for each key vehicle part detection using RPN.

Accuracy	bus	car	train	truck	tricar	backmirror	hungs	lightcover	windglass
%	**79.4**	**87.4**	**74.0**	**89.3**	25.9	**56.9**	0	24.9	**76.8**
Accuracy	anu signs	entrylicense	safebelt	plate	carlight	car top window	carrier	wheel	logo
%	9.1	2.3	1.8	**69.5**	**61.6**	**61.3**	21.4	**52.8**	**51.7**

Preliminary on RPN. To prove the reliability of RPN, we first train RPN on our extra dataset to generate interested region proposals. Fig. 3 depicts the original images in our dataset and the corresponding detection results using RPN, from which we can draw an intuitional conclusion that even under these challenging conditions our RPN can achieve satisfying performance. Meanwhile, Tab. 1 shows the detailed accuracy of detecting each key vehicle part by RPN. We found that several vehicle parts are nearly detected like hungs, anu signs, safebelt and so on, that is because these vehicle parts hardly appear when captured by surveillance cameras. Here we cut off the vehicle parts whose accuracies are not bold in Tab. 1. Therefore, high accuracy on localizing key vehicle parts makes it more reliable for extracting more distinguishable local features.

Impact of feature extraction module. The feature extraction module contains three critical components in our EIA-Net: RPN for locating the key vehicle parts, R-MAC descriptor for construction of local features on top of the detected key vehicle parts, and Part-based CNN for supplementing the local and global features. In order to verify the effects of these critical components, we derive

several variants to make comparisons with baselines. Using the same traditional classifiers, *i.e.*, SVM and Joint Bayesian, we derive two pairs of variants by replacing the standard CNN fully-connected features respectively with features extracted by RPN with or without R-MAC construction, named as "RPN + FC feature + SVM" / "RPN + FC feature + Joint Bayesian" and "RPN + R-MAC + SVM" / "RPN + R-MAC + Joint Bayesian". As mentioned before, the baseline "Plain DRDL" is actually the simplified overall network that includes two-branch CNN network trained by CCL. On top of this baseline, we derive two variants by adding one of the two components to further prove the effectiveness of them, *i.e.*, RPN and R-MAC, and Part-based CNN, hereinafter called "RPN + R-MAC + DRDL*" and "RPN + CR-MAC + DRDL*" (CR-MAC represents the feature concatenated by R-MAC and features extracted by Part-based CNN). Specifically, in "RPN + R-MAC + DRDL*" variant, we only employ the image representation of the last convolutional responses (R-MAC descriptor) given a set of detected region proposals while "RPN + CR-MAC + DRDL*" exploits the combined features concatenated with Part-based CNN fully-connected layer feature.

Table 2. Accuracy comparison of baselines and our derived variants for vehicle model verification on CompCars and VehicleID. The first three rows represent the baseline methods while the rest of rows denote our variants.

Method	CompCars			VehicleID
	Easy	Medium	Hard	
FC feature + SVM [25]	0.700	0.690	0.659	0.683
FC feature + Joint Bayesian [25]	0.833	0.824	0.761	0.810
Plain DRDL [58]	0.828	0.788	0.703	0.775
RPN + FC feature + SVM	0.742	0.728	0.684	0.721
RPN + FC feature + Joint Bayesian	0.859	0.845	0.782	0.836
RPN + R-MAC + SVM	0.765	0.743	0.702	0.738
RPN + R-MAC + Joint Bayesian	0.874	0.858	0.799	0.847
RPN + R-MAC + DRDL*	0.873	0.846	0.772	0.839
RPN + CR-MAC + DRDL*	0.875	0.859	0.787	0.848
RPN + CR-MAC + EIAL + DRDL*	0.882	0.869	0.803	**0.866**
RPN + CR-MAC + EIAL + MCCL	**0.885**	**0.872**	**0.814**	0.863

From Tab. 2, in contrast with the traditional classifier baselines (the first two rows), though we use the fully-connected layer feature, we obtain a 2% to 4% performance gain with SVM and around 2% with Joint Bayesian by merely taking advantage of RPN, which proves the superiority of locating relatively explicit key vehicle parts from the side. With the R-MAC aggregating the local features of detected key vehicle parts by RPN, we further boost the performance by around 2% both with SVM and Joint Bayesian. Similarly, our "RPN + R-MAC + DRDL*" variant outperforms the "Plain DRDL" baseline by around 4% to 6%. Note that the DRDL method is not specifically devised for vehicle model verification task, however, the experimental results of the "RPN + R-MAC + DRDL*" variant are comparable with those of the "RPN + R-MAC + Joint Bayesian" variant. That means that the influence of RPN and R-MAC can provide more effective information when the classifier or framework possess weaker recognizing capacity. Compared with "RPN + R-MAC + DRDL*", we observe that "RPN + CR-MAC + DRDL*" attain an improvement by around 1%. We analyze that R-MAC descriptors are actually the aggregation of local features, while Part-based CNN can assist to performance gain by supplementing for the global features from its full-body channel. We also conduct experiments with different numbers of channels and convolutional layers in Part-based CNN, finally we adopt the settings as follows: (1) The first convolutional layer of the overall network

is cut into two equal body-parts, which means Part-based CNN has two body-part channel and one global full-body channel; (2) Full-body channel contains two convolutional layers and two pooling layers; (3) Each body-part channel contains two convolutional layers. Above all, together with RPN, R-MAC, and Part-based CNN, the accuracy of our method has been promoted to about 2% to 3% higher than the-state-of-the-art "FC feature + Joint Bayesian".

Impact of deep metric learning module. On top of the aforementioned variant "RPN + CR-MAC + DRDL*", we also derive two variants concerning the deep metric learning module to further prove its effectiveness, named as "RPN + CR-MAC + EIAL + DRDL*" and "RPN + CR-MAC + EIAL + MCCL", corresponding to EIAL and modified CCL respectively. As can be seen in Tab. 2, we conclude that the proposed EIAL promotes the performance by around 1%. We suppose that EIAL can alleviate the common misalignment issue existing in verification or Re-ID tasks. An interesting phenomenon can also be observed from Tab. 2 that the modified CCL nearly improves the performance. That is our modified CCL is specially designed for finer distinguishing. Please refer to the Tab. 3 and Tab. 4, the performance of all the methods decrease when involving the model verification on the released year level, but our proposed method trained with modified CCL suffer far more less performance loss than other methods, which verify the critical role of modified CCL. To sum up, our complete EIA-Net with all of the crucial components outperforms the state-of-the-arts by at least 5 percent.

4.4 Performance Comparison for Vehicle Model Verification

Table 3. Accuracy comparison of our method and state-of-the-arts for model-level vehicle model verification on CompCars and VehicleID. * indicates the re-implemented methods. "_" indicate the second best results.

Method	CompCars			VehicleID
	Easy	Medium	Hard	
FC feature + SVM [25]	0.700	0.690	0.659	0.683
FC feature + Joint Bayesian [25]	0.833	0.824	0.761	0.810
DRDL [58]	0.828	0.788	0.703	0.775
PRN* [14]	0.863	0.842	0.795	0.848
PVEN* [34]	**0.894**	0.865	<u>0.807</u>	0.854
VehicleNet* [64]	0.878	0.862	0.801	0.849
Ours with CCL	0.882	<u>0.869</u>	0.803	**0.866**
Ours with MCCL	<u>0.885</u>	**0.872**	**0.814**	<u>0.863</u>

In this subsection, we carry out the experiments using the well-trained EIA-Net including a complete set of crucial components on CompCars and VehicleID. To validate the effectiveness of EIA-Net, we perform accuracy comparison for vehicle model verification on both the make-level and the released-year-level, where the former represents a more corse recognition while the latter represents a finer one. To ensure the experimental fairness, we also re-implement several vehicle Re-ID approaches by feeding model labels for comparisons, which are indicated by * in Tab. 3 and Tab. 4. With the goal of further investigating the local feature extraction of EIA-Net, the most discriminative regions of several example samples are illustrated in the form of heatmap.

Performance comparison on CompCars and VehicleID. Tab. 3 and Tab. 4 depict the performance comparisons on the the make-level and the released-year-level, respectively. In Tab. 3 and Tab. 4, we can see that our method with the original version of CCL exceeds the state-of-the-arts and Re-ID methods under three dataset settings, and our method achieves the best under all dataset

Table 4. Accuracy comparison of our method and state-of-the-arts for released-year-level vehicle model verification on CompCars and VehicleID. * indicates the re-implemented methods. "_" indicate the second best results.

Method	CompCars			VehicleID
	Easy	Medium	Hard	
FC feature + SVM [25]	0.683	0.671	0.636	0.668
FC feature + Joint Bayesian [25]	0.819	0.810	0.745	0.794
DRDL [58]	0.813	0.774	0.691	0.766
PRN* [14]	0.847	0.835	0.768	0.814
PVEN* [34]	0.857	0.845	<u>0.793</u>	0.836
VehicleNet* [64]	<u>0.869</u>	<u>0.851</u>	0.781	0.835
Ours with CCL	0.865	0.847	0.782	<u>0.838</u>
Ours with MCCL	**0.879**	**0.864**	**0.806**	**0.851**

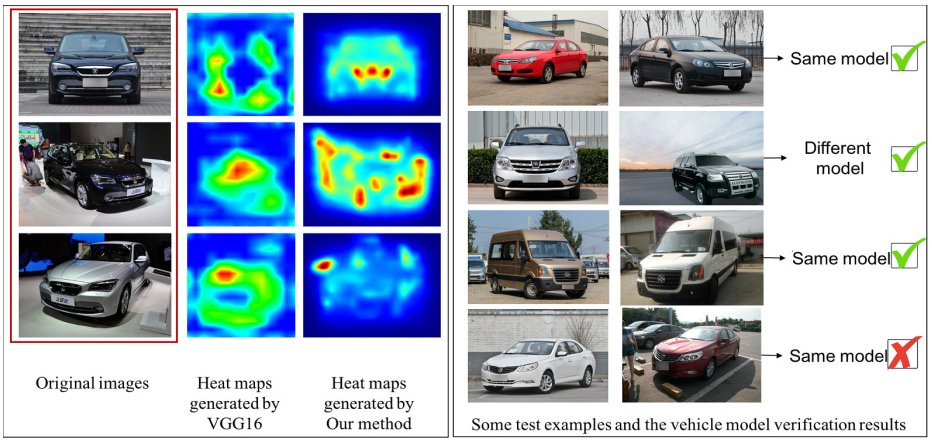

Fig. 4. Left: The illustration of several sample images and their corresponding heat maps of plain VGG16 and the RPN in EIA-Net. Right: The illustration of several test examples for vehicle model verification and their prediction results.

settings. Our method is superior than those Re-ID methods with bounding boxes annotation (like PRN) and even finer labeling (like PVEN), since ours also discovers discriminative features to represent vehicle variance and Re-ID methods are not specifically designed for model verification and thus less competitive. From Tab. 4, we notice that there is a minor decrease on accuracy in contrast to the experimental results that ignore the released year. We observe that most errors lie in the image pairs with different released years but of the same vehicle model due to the natural inter-class similarity issue, *i.e.*, the vehicle make usually changes a little on the appearance of certain parts and the varieties of viewpoints can cause the occlusion of the key vehicle parts. However, as can be seen in Tab. 4, in contrast with Tab. 3, our method with modified CCL achieves a much larger performance gain than the state-of-the-arts and our method with original CCL, which proves the effectiveness of the modified CCL in improving the finer recognition for vehicle models.

Further discussion. We perform a supplementary experiment that visualizes the ReLU_5 layer feature maps as heat maps, and the visualization tool we used here is [39]. Fig. 4 (Left) indicates the original images, heat maps generated by plain VGG16 [44] and our adopted RPN from the left column to the right one. We can clearly notice that the activations of VGG16 spread across the whole image, while the activations of RPN are constricted to several interested regions, *i.e.*, key vehicle parts. However, this phenomenon is not obvious in the third heat map, that is because sometimes the performance of RPN is affected by various circumstances such as illumination and change of viewpoints. In addition, some test examples for vehicle model verification and their prediction results are illustrated in Fig. 4 (Right). Our proposed method can distinguish different models with subtle appearance differences, or the same models with large color difference. Nevertheless, our proposed approach gives the wrong predictions for the last image pairs. These two vehicle images come from the same make and model, but they belong to the different released years. Considering this situation, it points out the direction of later improvement of our approach.

4.5 Generalization to Vehicle Re-identification

Table 5. Comparison with the state-of-the-arts in terms of Rank-1(%) and mAP (%) on the VeRi-776 dataset and Rank-1(%) and Rank-5 (%) on the VehicleID dataset. "_" indicate the second best results.

Methods	VeRi-776		VehicleID(Small)		VehicleID(Medium)		VehicleID(Large)	
	Rank-1	mAP	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
LOMO [23]	23.87	9.78	19.74	32.14	18.95	29.46	15.26	25.63
GoogLeNet [58]	52.12	17.81	47.90	67.43	43.45	63.53	38.24	59.51
FACT [26]	51.85	18.73	49.53	67.96	44.63	64.19	39.91	60.49
XVGAN [66]	60.20	24.65	52.89	80.84	-	-	-	-
SiameseCNN [41]	41.12	29.48	-	-	-	-	-	-
OIFE [53]	65.92	48.00	-	-	-	-	67.0	82.9
VAMI [67]	77.03	50.13	63.12	83.25	52.87	75.12	47.34	70.29
NuFACT [27]	81.56	53.42	48.90	69.51	43.64	65.34	38.63	60.72
AAVER [19]	88.68	58.52	72.47	93.22	66.85	89.39	60.23	84.85
VANet [68]	89.78	66.34	83.26	95.97	81.11	94.71	77.21	92.92
PAMTRI [50]	92.86	71.88	-	-	-	-	-	-
SAN [35]	93.3	72.5	79.7	94.3	78.4	91.3	75.6	88.3
PRN [14]	94.3	74.3	78.4	92.3	75.0	88.3	74.2	86.4
PVEN[34]	95.6	79.5	**84.7**	**97.0**	80.6	94.5	77.8	92.0
VehicleNet[64]	**96.78**	**83.41**	83.64	96.86	81.35	93.61	**79.46**	92.04
Ours	95.69	79.32	84.13	96.47	**81.92**	**94.88**	79.15	**93.36**

Table 6. Comparison with the state-of-the-arts in terms of Rank-1(%) and mAP (%) on the VERI-Wild dataset. "_" indicate the second best results.

Methods	VERI-Wild(Small)		VERI-Wild(Medium)		VERI-Wild(Large)	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
GoogLeNet [58]	24.3	57.2	24.2	53.2	21.5	44.6
DRDL [58]	22.5	57.0	19.3	51.9	14.8	44.6
FDA-Net [32]	35.1	64.0	29.8	57.8	22.8	49.4
MLSL [1]	46.3	86.0	42.4	83.0	36.6	77.5
PCRNet [28]	81.2	92.5	75.3	89.6	67.1	85.0
PVEN [34]	**82.5**	**96.7**	77.0	95.4	**69.7**	93.4
Ours	81.67	95.34	**78.15**	**95.63**	68.26	**93.82**

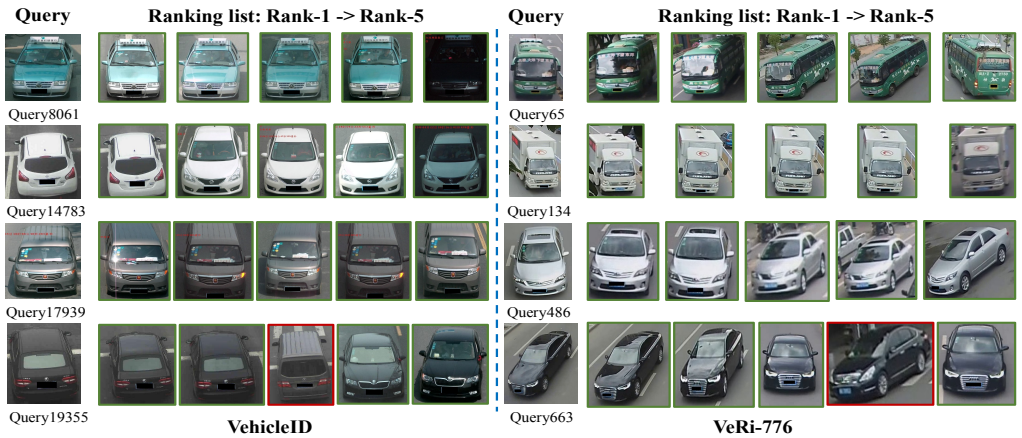

Fig. 5. The visualization of ranking list for vehicle Re-ID task on VehicleID and VeRi-776. The images in the first column denote the query images and the rest of images are top-5 most similar ranking results. The correct retrieved images are depicted in green bounding boxes, while the false instances are depicted in red ones.

To validate the robustness and generalization capacity of our proposed EIA-Net, we further perform comprehensive experiments using the complete EIA-Net for the vehicle Re-ID task on two popular datasets, *i.e.*, VehicleID and VeRi-776. To perform a fair comparison, we make a slight change in EIA-Net by replacing the backbone with ResNet-50 [15] instead of VGG16. Following the setting in [14], we resize the input image to 256×256. During the test stage, we did not adopt other post-processing strategies regardless of applying the mean feature of images flipped horizontally.

Comparison with the state-of-the-arts on VeRi-776 and VehicleID. Tab. 5 lists the experimental comparison results on VeRi-776 and VehicleID dataset, adopting Rank-1/mAP and Rank-1/Rank-5 evaluation metrics, respectively. By comparison with the state-of-the-arts, our proposed method exceeds those methods without finer labeling such as segmentation masks or domain transfer learning by a relatively large margin. Compared with recent works which introduce finer semantic information, *i.e.*, PVEN with segmentation masks and VehicleNet with domain prior learned from other vehicle Re-ID datasets, our method normally achieves the best two performance under most dataset settings. Specifically, our method shows the best performance in Medium subset of VehicleID. From Tab. 5, we can also observe that recent methods focusing on the fusion of local features and global features has become a trend in this task. Indeed, those approaches with emphasis on extracting local visual cues tend to possess more powerful recognition capability than others, which demonstrates again that local features are more important than the global ones. For example, the mAP has increased by more than 10% even 20%, while the Rank-1 is also boosted from less than 80% to around 90%. Nowadays, among the existing local feature based methods, state-of-the-arts are normally those that introduce extra data with annotations, such as bounding boxes in PRN, keypoints, viewpoints and pixel labeling. The experimental results demonstrate that methods with bounding boxes perform better than those with keypoints and viewpoints and remain competitive with those with finer annotation. That is regional prior, *i.e.*, key vehicle parts, can enforce the model pay more attention to crucial local regions, *e.g.*, logos, head lights, even the personalized stuff in the windshield, to further alleviate the misalignment issues. Concerning resolving the misalignment issue, our proposed method is still

superior than other bounding box related methods like PRN. We assume that one reason lies in the EIAL that puts stronger constraints on the key vehicle parts.

Comparison with the state-of-the-arts on VERI-Wild. Tab. 6 illustrates the experimental results on VERI-Wild dataset. From Tab. 6, we can observe the same phenomenon as Tab. 5. That is our method exceeds those approaches without finer labelling, *i.e.*, vehicle part parsing information, and shows competitive performance compared to recent methods with pixel semantics such as PVEN and PCRNet. Particularly, our method still achieves the best performance on Medium subset of VERI-Wild.

Visualization of Vehicle Re-ID Results. As shown in Fig. 5, we provide the qualitative image search results on two benchmarks, *i.e.*, VehicleID and VeRi-776. For each dataset, four query images and corresponding ranking lists are illustrated, from which we can observe that the retrieved results of our proposed method are rather satisfying, with the top-5 of the ranking list are nearly the relevant images.

5 CONCLUSION

In this paper, we investigate the problem of vehicle model verification that has not been explored deeply before. Unlike the previous works, we go beyond the vehicle model verification to vehicle model verification of different released years, such as "MINI CLUBMAN" 2008, 2009, 2011, 2013 and 2014. There naturally exist subtle differences between those models of different released years, which might be reflected by the change of some crucial vehicle parts such as car light. However, the former researches solve the task by either referring to the experiences on Face Verification or existing deep learning method. Both of them adopt the fully-connected layer features as global image descriptors, which can not better represent the local key part information of the vehicles.

To address the issues mentioned above, we propose an EIA-Net which is comprised of two modules: feature extraction module and deep metric learning module. The former includes an RPN-based network that aggregates the local features of key vehicle parts to form a compact feature vector with more discriminative abilities, and a Part-based CNN that supplements global representations for the final image features. The latter indicates two loss functions, *i.e.*, enforced intra-class alignment loss and modified coupled cluster loss, to learn an embedding into a latent distance space. Specifically, the RPN-based network is trained on our vehicle dataset with extra bounding boxes as annotations and used for generating region proposals that localize the key vehicle parts such as logo, plate, car light *etc.* To further prove the robustness of EIA-Net, we generalize the model to the vehicle Re-ID task. Extensive experimental results have demonstrated the effectiveness of EIA-Net compared with the-state-of-the-arts for both the vehicle model verification task and vehicle Re-ID task.

6 ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62072027, Grant 61872032, and 62076021; in part by the Beijing Natural Science Foundation under Grant 4202057, Grant 4202058, and Grant 4202060.

REFERENCES

[1] S. Alfasly, Y. Hu, H. Li, T. Liang, X. Jin, B. Liu, and Q. Zhao. 2019. Multi-label-based similarity learning for vehicle re-identification. *IEEE Access.* 7, 16 (2019), 2605–2616.

[2] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan. 2018. Group-Sensitive Triplet Embedding for Vehicle Re-identification. *IEEE Trans. Multimedia.* 20, 9 (2018), 2385–2399.

[3] D. Chen, X. Cao, L. Wang, and F. Wen. 2012. Bayesian face revisited: A joint formulation.. In *Proc. Eur. Conf. Comp. Vis.*

[4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 21,

1 (2016), 1âĂŞ13.
[5] W. Chen, X. Chen, J. Zhang, and K. Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. 2016. Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[7] S. Chopra, R. Hadsell, and Y. Lecun. 2005. Learning a similarity metric discriminatively,with application to face verification.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[8] M. Cormier, L. W. Sommer, and M. Teutsch. 2016. Low resolution vehicle re-identification based on appearance features for wide area motion imagery.. In *Proc. IEEE Winter Appl. Comput. Vis. Workshops.*
[9] P. Cui, S. Liu, and W. Zhu. 2018. General knowledge embedded image representation learning. *IEEE Trans. Multimedia.* 20, 1 (2018), 198–207.
[10] A. Babenko et al. 2015. Aggregating Deep Convolutional Features for Image Retrieval.. In *Proc. IEEE Int. Conf. Comp. Vis.*
[11] R. S. Feris, B. Siddiquie, J. Petterson, Y. Zhai, A. Datta, and L. M. Brown. 2012. Group-Sensitive Triplet Embedding for Vehicle Re-identification. *IEEE Trans. Multimedia.* 14, 1 (2012), 28–42.
[12] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. 2016. Deep Image Retrieval: Learning Global Representations for Image Search.. In *Proc. Eur. Conf. Comp. Vis.*
[13] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu. 2018. Learning coarse-to-fine structured feature embedding for vehicle re-identification.. In *Proc. AAAI Conf. Art. Int.*
[14] B. He, J. Li, Y. Zhao, and Y. Tian. 2019. Part-regularized near-duplicate vehicle re-identification.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[15] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[16] A. Hossein, R. A. Sharif, S. Josephine, M. Atsuto, and C. Stefan. 2014. From generic to specific deep representations for visual recognition. *In arXiv preprint* (2014).
[17] E. Hsiao, S. N. Sinha, K. Ramnath, S. Baker, L. Zitnick, and R. Szeliski. 2014. Car make and model recognition using 3d curve alignment.. In *Proc. IEEE Winter Appl. Comput. Vis.*
[18] J. Hu, J. Lu, and Y. P. Tan. 2014. Discriminative deep metric learning for face verification in the wild.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[19] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J. Chen, and R. Chellappa. 2019. A dual-path model with adaptive attention for vehicle re-identification.. In *Proc. IEEE Int. Conf. Comp. Vis.*
[20] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 2013. 3d Object representations for fine-grained categorization.. In *Proc. IEEE Int. Conf. Comp. Vis. Workshops*.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks.. In *Proc. Adv. Neural Inf. Process. Syst.*
[22] Z. Li and J. Tang. 2015. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Trans. Multimedia.* 17, 11 (2015), 1989–1999.
[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. 2015. Person re-identification by local maximal occurrence representation and metric learning.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[24] Y.-L. Lin, V. I. Morariun, W. Hsu, and L. S. Davis. 2014. Jointly optimizing 3D model fitting and fine-grained classification.. In *Proc. Eur. Conf. Comp. Vis.*
[25] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. 2016. Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[26] X. Liu, W. Liu, T. Mei, and H. Ma. 2016. A deep learning-based approach to progressive vehicle re-identification for urban surveillance.. In *Proc. Eur. Conf. Comp. Vis.*
[27] X. Liu, W. Liu, T. Mei, and H. Ma. 2018. PROVID: progressive and multimodal vehicle re-identification for large-scale urban surveillance. *IEEE Trans. Multimedia.* 20, 3 (2018), 645–658.
[28] X. Liu, W. Liu, J. Zheng, C. Yan, and Tao Mei. 2020. Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification. In *Proc. ACM Multimedia.*
[29] X. Liu, S. Zhang, Q. Huang, and W. Gao. 2018. RAM: A region-aware deep model for vehicle re-identification.. In *Proc. IEEE Int. Conf. Multimedia Expo.*
[30] J. Long, E. Shelhamer, and T. Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*
[31] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan. 2019. Embedding adversarial learning for vehicle re-identification. *IEEE Trans. Image Proc.* 28, 8 (2019), 3794–3807.
[32] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[33] X. Ma, T. Zhang, and C. Xu. 2019. GCAN: Graph Convolutional Adversarial Network for Unsupervised Domain Adaptation.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[34] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z. Zha, X. Gao, S. Wang, and Q. Huang. 2020. Parsing-based View-aware Embedding Network for Vehicle Re-Identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[35] J. Qian, W. Jiang, H. Luo, and H. Yu. 2019. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification.. In *arXiv:1910.05549.*

[36] K. Ramnath, S. N. Sinha, R. Szeliski, and E. Hsiao. 2014. Car make and model recognition using 3d curve alignment. *In: Applications of Computer Vision (WACV)* (2014).

[37] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.. In *Proc. Adv. Neural Inf. Process. Syst.*

[38] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *In arXiv preprint arXiv:1610.02391* (2017).

[40] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *In arXiv preprint arXiv:1312.6229* (2013).

[41] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. 2017. Learning deep neural networks for vehicle Re-ID with visual-spatio-temporal path proposals.. In *Proc. IEEE Int. Conf. Comp. Vis.*

[42] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. 2015. Discriminative learning of deep convolutional feature point descriptors.. In *Proc. IEEE Int. Conf. Comp. Vis.*

[43] K. Simonyan and A. Zisserman. 2012. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2012).

[44] K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.. In *Int. Conf. Lea. Repr.*

[45] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. 2016. Deep metric learning via lifted structured feature embedding.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[46] Y. Sun, Y. Chen, and X. Wang. 2014. Deep Learning Face Representation by Joint Identification-Verification.. In *Proc. Adv. Neural Inf. Process. Syst.*

[47] Y. Sun, X. Wang, and X. Tang. 2014. Deep learning face representation from predicting 10,000 classes.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[48] Y. Sun, X. Wang, and X. Tang. 2015. Deeply learned face representations are sparse, selective, and robust.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[49] Y. Taigman, M. Yang, M. Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[50] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang. 2019. PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data.. In *Proc. IEEE Int. Conf. Comp. Vis.*

[51] G. Tolias, R. Sicre, and H. Jegou. 2016. Particular Object Retrieval With Integral Max-Pooling of CNN Activations.. In *Int. Conf. Lea. Repr.*

[52] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. 2014. Learning fine-grained image similarity with deep ranking.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[53] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang. 2017. Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification.. In *Proc. IEEE Int. Conf. Comp. Vis.*

[54] K. Q. Weinberger and L. K. Saul. 2015. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10, 1 (2015), 207–224.

[55] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. 2015. Data-driven 3d voxel patterns for object category recognition.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[56] S. Xie, Z. Zheng, L. Chen, and C. Chen. 2018. Learning semantic representations for unsupervised domain adaptation.. In *Proc. Int. Conf. Mach. Learn.*

[57] X.Yang, M.Wang, and D. Tao. 2018. Person re-identification with metric learning using privileged information. *IEEE Trans. Image Proc.* 27, 2 (2018), 791–805.

[58] L. Yang, P. Luo, C. C. Loy, and X. Tang. 2015. A large-scale Car Dataset for Fine-grained Categorization and Verification.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[59] X. Yang, M. Wang, R. Hong, Q. Tian, and Y. Rui. 2017. Enhancing person re-identification in a self-trained subspace. *Proc. ACM Trans. Multimed. Comput. Commun. Appl.* 13, 3 (2017), 1âĂŞ23.

[60] D. Yi, Z. Lei, S. Liao, and S. Z. Li. 2014. Deep Metric Learning for Person Re-Identification.. In *Proc. IEEE Int. Conf. Patt. Recogn.*

[61] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. 2016. Embedding label structures for fine-grained feature representation.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[62] Z. Zhang, T. Tan, K. Huang, and Y. Wang. 2012. Three dimensional deformable-model-based localization and recognition of road vehicles. *IEEE Trans. Image Proc.* 21, 1 (2012), 1âĂŞ13.

[63] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. 2015. Scalable Person Re-identification: A Benchmark.. In *Proc. IEEE Int. Conf. Comp. Vis.*

[64] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei. 2020. VehicleNet: Learning Robust Visual Representation for Vehicle Re-identification. *IEEE Trans. Multimedia.* (2020).

[65] F. Zhou and Y. Lin. 2016. Fine-grained image classification by exploring bipartite-graph labels.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[66] Y. Zhou and L. Shao. 2017. Cross-view gan based vehicle generation for re-identification.. In *Proc. Brit. Mach. Vis. Conf.*

[67] Y. Zhou and L. Shao. 2018. Aware attentive multi-view inference for vehicle re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[68] Y. Zhou and L. Shao. 2018. Viewpoint-aware attentive multi-view inference for vehicle re-Identification.. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

[69] J. Zhu, X. Chen, and A. L. Yuille. 2016. DeepM: A Deep Part-Based model for object detection and semantic part localization.. In *Int. Conf. Lea. Repr.*