Disentangled Feature Learning Network for Vehicle Re-Identification

Yan Bai[1,3], **Yihang Lou**[1,3], **Yongxing Dai**[1,3], **Jun Liu**[2], **Ziqian Chen**[1,3] and **Ling-Yu Duan**[1,3*]

[1] National Engineering Lab for Video Technology, Peking University, Beijing, China
[2] ISTD Pillar, Singapore University of Technology and Design, Singapore
[3] Peng Cheng Laboratory, Shenzhen, China

{yanbai, yihanglou, yongxingdai}@pku.edu.cn, jun_liu@sutd.edu.sg, {wzziqian,lingyu}@pku.edu.cn

Abstract

Vehicle Re-Identification (ReID) has attracted lots of research efforts due to its great significance to the public security. In vehicle ReID, we aim to learn features that are powerful in discriminating subtle differences between vehicles which are visually similar, and also robust against different orientations of the same vehicle. However, these two characteristics are hard to be encapsulated into a single feature representation simultaneously with unified supervision. Here we propose a Disentangled Feature Learning Network (DFLNet) to learn orientation specific and common features concurrently, which are discriminative at details and invariant to orientations, respectively. Moreover, to effectively use these two types of features for ReID, we further design a feature metric alignment scheme to ensure the consistency of the metric scales. The experiments show the effectiveness of our method that achieves state-of-the-art performance on three challenging datasets.

1 Introduction

Vehicle Re-Identification (ReID) aims to retrieve all the images of a given query vehicle identity, from a large image database. Deep learning techniques have greatly promoted the development of vehicle ReID in the past few years. Many previous works [Bulan *et al.*, 2017] conduct vehicle ReID as a license plate recognition procedure. However, license plate recognition requires high-resolution images captured with front or rear views. Moreover, in some extreme cases, the license plates may be deliberately removed, occluded, or even faked. Recent works [Liu *et al.*, 2016b; Liu *et al.*, 2016c] start to focus on visual feature-based ReID, where ReID is performed as feature matching between the query and reference vehicle images.

Orientation is a crucial factor in vehicle ReID. Given two vehicle images captured from the same orientation, we expect the extracted features are capable of encoding vehicle subtle details, such as the tissue boxes or air inlet of engine, as illustrated in Fig.1. These details are orientation specific information. The works in [He *et al.*, 2019; Pirazh *et al.*, 2019] use local part and keypoint information to learn orientation specific features and show the efficacy of such features for vehicle ReID. In contrast, when given two vehicle images with different orientations, the learned features are desired to capture orientation-invariant common characteristics of the vehicles, such as colors and vehicle design styles. The methods using orientation invariant feature embedding [Wang *et al.*, 2017] and viewpoint-aware metrics [Chu *et al.*, 2019] show the benefits of such orientation invariant information for cross-view ReID. Though orientation specific and orientation common information are both useful for vehicle ReID, the previous methods often focus on learning a single representation which is however difficult to simultaneously capture these two types of powerful information.

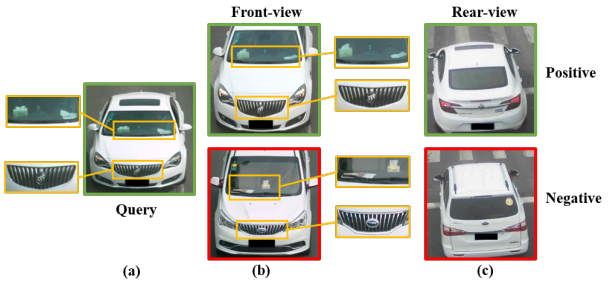

Figure 1: Illustration of vehicle matching. For the same orientation ((a) with (b)), the specific details are important cues for matching. For different orientations ((a) with (c)), the details may become useless, while the interior common characteristics are very important.

Actually, such two characteristics are hard to be simultaneously obtained with a single feature representation learned under unified supervision. Moreover, using a single feature representation, it becomes difficult to flexibly choose the proper information for vehicle matching under variant conditions. For example, the vehicle details such as decorations in vehicle front-view are important for retrieving front-view vehicles, but may even degrade the performances of retrieving the rear-view ones, as shown in Fig. 1. This means when comparing vehicle pairs under the same orientation, orientation specific features are often useful, while the orientation common features are important to recall the same vehicles with huge orientation variations.

*Corresponding Author

This motivates us to disentangle the learning of these two types of features from a single embedding model and design an adaptive matching method for ReID. Concretely, instead of learning a single feature representation, we propose a Disentangled Feature Learning Network (DFLNet) to explicitly learn the orientation specific and orientation common features concurrently, i.e., the learning of the two types of information is disentangled yet within a joint network.

Moreover, to learn powerful common features for ReID under variant orientations, we propose a novel "Odd-One-Out" adversarial scheme that distills the orientation invariant information shared by all samples of the same vehicles to obtain strong orientation independent ability. Besides, to learn orientation specific features, we also design an attention scheme to mine and focus on the useful detailed information.

To effectively utilize these two types of features, in this paper, a hybrid ranking strategy with feature metric alignment is also designed for adaptive matching in the ReID procedure.

Overall, our contributions can be summarized as follows:

- A novel DFLNet is proposed to explicitly learn orientation common and specific features jointly for Vehicle ReID.
- A novel "Odd-One-Out" adversarial scheme is proposed to learn representative common features. An attention module is also designed for specific feature learning.
- A hybrid ranking strategy is designed to take advantages of specific and common features in the ReID procedure.
- Our DFLNet achieves state-of-the-art performances on all the evaluated benchmarks. It brings 12% mAP gains on cross-view ReID compared to the baseline model.

2 Related Work

Vehicle ReID. Recently, vehicle ReID has attracted much research focus [Liu *et al.*, 2016a; He *et al.*, 2019]. Some methods [He *et al.*, 2019] focus on improving the discriminative capability of the models to distinguish specific subtle details of similar vehicles. Liu *et al.* [Liu *et al.*, 2016a] introduce a mixed difference network in which both the vehicle model and ID information are used as supervisions for learning an embedding model. He *et al.* [He *et al.*, 2019] propose to use vehicle part information to regularize the global feature learning. In [Pirazh *et al.*, 2019], Pirazh *et al,* propose a dual-path model with adaptive attention model, which is able to get orientation conditioned keypoints to extract local features.

Some other methods focus on learning orientation invariant features. In [Wang *et al.*, 2017], Wang *et al.* propose to use vehicle keypoint localization to align and generate orientation invariant features for vehicle ReID. In [Zhou and Shao, 2018], Zhou *et al.* propose a multi-view feature inference scheme, which uses a single-view input vehicle to generate the multi-view features. In [Tang *et al.*, 2019], Tang *et al.* propose a pose-aware multi-task learning method using synthetic data to learn viewpoint invariant features.

A few methods have considered using these two types of information simultaneously. In [Chu *et al.*, 2019], Chu *et al.* use two triplets constraints for the same and different viewpoints in two feature spaces. Bai *et al.* [Bai *et al.*, 2018] propose a group sensitive triplet embedding model to build up a type of "similar attribute, closer distance" feature embedding by a two-level margin constraint. However, they both learn and represent these two types of information under the same optimization objectives, and the differences between these two types of features have not been explicitly investigated.

Different from all the aforementioned works, we propose a disentangled feature learning network to learn these two types of features concurrently. The common features are the invariant and consistent representations shared by all samples of the same vehicles, and the specific features are the representations exploiting the subtle difference cues.

Disentangled Representation. Disentangled schemes have been used in image generation [Ma *et al.*, 2017] and pose-invariant representation learning [Tran *et al.*, 2017]. Tran *et al.* [Tran *et al.*, 2017] propose explicit disentangled representations based on face variations through pose codes. Zhao *et al.* [Zhao *et al.*, 2019] propose an attribute-driven method to disentangle several sub-features corresponding to semantic attribute groups for video-based person ReID. However, these methods disentangle sub-features based on each separate attribute, and common representations are ignored.

3 Proposed Method

The architecture of our proposed DFLNet is illustrated in Fig. 2. In training stage, the orientation common features are learned by adversarial learning with an "Odd-One-Out" scheme. The orientation specific features are learned by an attention scheme in a triplet embedding design. During testing, we design a hybrid ranking strategy with a feature metric alignment scheme. We use common features to get initial recall list and use specific features to compare recall samples with the same orientation as the query for re-ranking.

3.1 Orientation Common Feature Learning by Odd-One-Out Adversary

To improve feature robustness under variant orientations, the ideal orientation common features are expected to encode the information shared by all samples of the same vehicle and ignore the orientation specific information, as the orientation specific information is often unuseful for cross-view ReID and may even degrade the performance. This means the common feature representation needs to be orientation independent. However, features learned with general embedding network designs [Schroff *et al.*, 2015a] often contain orientation specific information. To get ideal common features, here we disentangle the common features from the base embedding features and distill the orientation invariant information.

Specifically, we design a novel "Odd-One-Out" adversarial learning scheme to generate orientation common features. A unit consisting of some vehicle images captured from the same orientation, together with one image captured from another orientation is constructed, and the "Odd-One" here means the sample has a different orientation compared to other samples in this unit. Concretely, we build a unit $<x^i, x^j, x^k>$ composed of three image samples of the same vehicle ID. $O(x^i) = O(x^j)$ and $O(x^i) \neq O(x^k)$, where

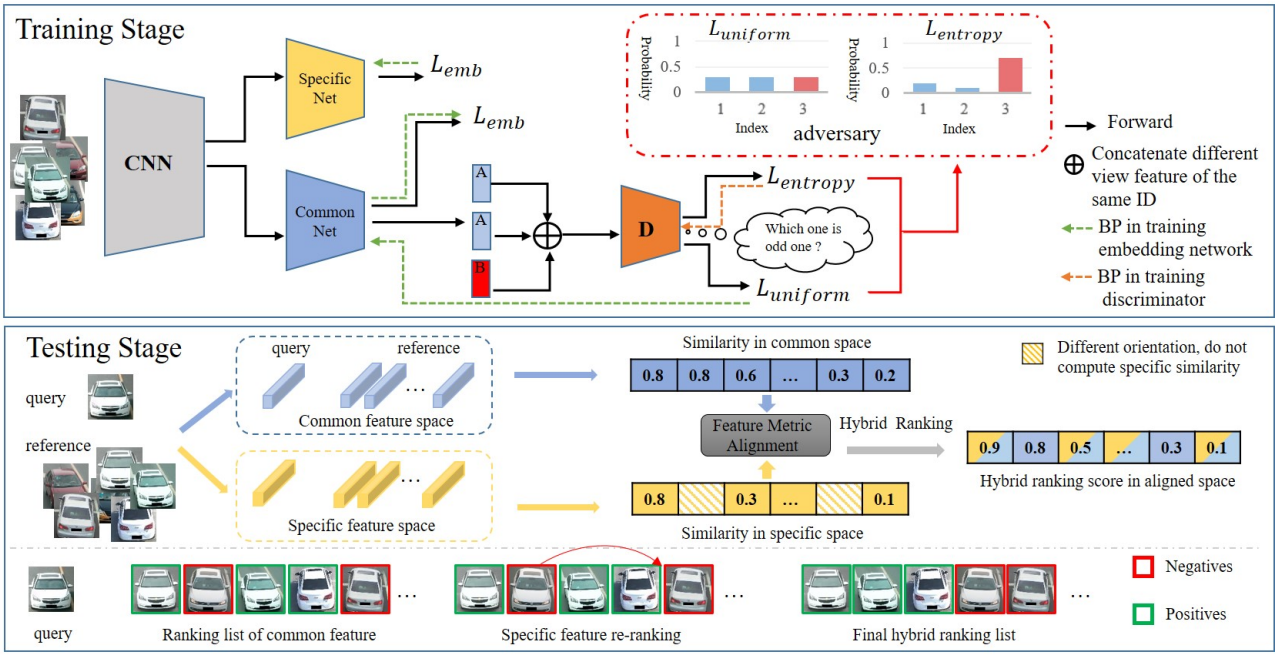

Figure 2: Illustration of our DFLNet framework. In the training stage, the DFLNet learns specific features and common features by the similarity constraint and the adversarial constraint, respectively. In the testing stage, we design a feature metric alignment mechanism for these two features, and a hybrid ranking is performed to get the final results.

$O(x^i)$ denotes the orientation of the sample x^i. In this case, sample x^k is the odd-one sample.

In our "Odd-One-Out" adversarial learning scheme, the discriminator D is pushed to learn and recognize which one is the odd-one sample, while the feature generator G aims to generate the common features that cannot be recognized by D. With the adversarial training going on, G will gain the ability of generating orientation independent features, i.e., the common features.

Common Feature Discriminator. When training the discriminator D, the generator G serves as an orientation common feature extractor. Given a vehicle sample unit $<x^i, x^j, x^k>$ from the same ID, we extract and concatenate their common features $G(.)$ as the input of D, *i.e.,* $[G(x^i); G(x^j); G(x^k)]$. D learns to predict which sample in the unit is the odd-one, and the position of the odd-one sample in the unit is used as the label for prediction. For example, for the unit $<x^i, x^j, x^k>$, if $O(x^i) = A$, $O(x^j) = A$, and $O(x^k) = B$, then the label of this unit is 3. This odd-one prediction process can be easily implemented as a classification task by using the cross entropy loss for supervision.

Common Feature Generator. When training the common feature generator G, the parameters in D are fixed. The ideal common features are independent to the variant orientations. Thus, we design a novel uniform loss $\mathcal{L}_{uniform}$, which constrains G to generate common features that will make D produce a uniform probability distribution on the odd-one prediction (classification). Such a constraint drives G to exploit the orientation independent information and try to eliminate the specific information in the generated features.

The $\mathcal{L}_{uniform}$ is inspired by label smoothing regularization [Szegedy *et al.*, 2016] that assigns small values to the non-groundtruth classes in cross entropy loss.

The standard cross entropy loss is formulated as:

$$\mathcal{L}_{entropy} = -\sum_{k=1}^{K} y_k log(\hat{y_k}), \tag{1}$$

where y_k is the ground-truth class label vector in one-hot distribution. $\hat{y_k}$ is the prediction probability of the input belonging to class k. For $\mathcal{L}_{uniform}$ loss, to be constant over all classes, the ground-truth class label distribution y_k in Eq. 1 is defined as $y_k = \frac{1}{K}$. Thus $\mathcal{L}_{uniform}$ loss is formulated as:

$$\begin{aligned} \mathcal{L}_{uniform} &= -\frac{1}{K}\sum_{k=1}^{K} log(\hat{y_k}), \\ \hat{y_k} &= \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \end{aligned} \tag{2}$$

where $\hat{y_k}$ is derived from the softmax function, and e^{z_k} is the output of the last fully connected layer in D.

In this manner, the adversarial relation between G (with $L_{uniform}$) and D (with $L_{entropy}$) can be constructed. G attempts to fool D to get a wrong odd-one prediction. Conversely, D tries to exploit the orientation specific information to predict which one is odd. As the adversarial learning goes on, in order to cheat D, G will try to eliminate the orientation specific information in the common features. Finally, only the orientation invariant information is reserved in the common features.

Note that since the common representation still needs to be discriminative for ReID, the cross entropy loss (for identity classification) and triplet loss (for distance metric learning) that will be introduced in Sec. 3.2 are both used during training the common feature learning modules.

3.2 Orientation Specific Feature Learning with an Attention Scheme

The orientation specific features need to exploit and encode subtle details of the vehicle images. It is beneficial to selectively focus on the informative regions that are useful for ReID, as shown by the yellow boxes in Fig. 1.

This selectively focusing scheme is also termed as attention that has been demonstrated to be effective in various areas, such as machine translation [Bahdanau *et al.*, 2014] and image caption generation [Xu *et al.*, 2015]. Therefore, we adopt the visual attention scheme to enable our network to learn and find the crucial details that need to be focused on.

In our method, the attention module computes the importance scores for each patch in the feature maps. Let x denotes the input and $f(x)$ denotes the representation obtained after network mapping. Meanwhile, the attention score $s(x)$ is generated by the attention module, which serves as gates for the base branch $b(x)$, as follows:

$$b_{i,j}(x) = f_{i,j}(x) \odot s_{i,j}(x), \tag{3}$$

where (i, j) indicates the patch position over the feature maps. This element-wise product can promote the responses on interested regions.

For learning orientation specific features, the triplet loss [Schroff *et al.*, 2015b] and cross-entropy loss are used for metric learning as below:

$$\mathcal{L}_{emb} = \omega \mathcal{L}_{entropy} + (1 - \omega)\mathcal{L}_{triplet}, \tag{4}$$

where ω is the weight in optimization. Using this scheme with the attention module, our network is thus able to learn discriminative orientation specific features.

3.3 Hybrid Ranking Strategy

To effectively take advantages of both the orientation common features and specific features, we propose a hybrid ranking strategy for vehicle ReID, in which the common and specific features are adaptively used for distance computation. Concretely, given a query image, we use common features to get initial recall list, then use specific features to compare samples with the same orientation as the query. Then, the hybrid ranking procedure can be formulated as:

$$d(x^i, x^j) = \begin{cases} d_c & if\ O(x^i) \neq O(x^j), \\ \lambda d_c + (1 - \lambda) f_{align}(d_s) & otherwise, \end{cases} \tag{5}$$

where d_c and d_s are the common and specific feature distances, respectively. The sample's orientation $O(\cdot)$ is obtained by training an orientation classification model. As orientation classification is a very simple task, we empirically observe the orientation can be recognized very accurately, *e.g.*, using MobileNet V2 [Sandler *et al.*, 2018] can achieve 99.5% classification accuracy on VehicleID dataset.

Note f_{align} in Eq. 5 is the feature distance mapping from specific feature to common feature, as we can not simply compare these two feature distances, since they are learned in different feature spaces. The feature distributions are shown in Fig. 3. The different distributions indicate the evaluation

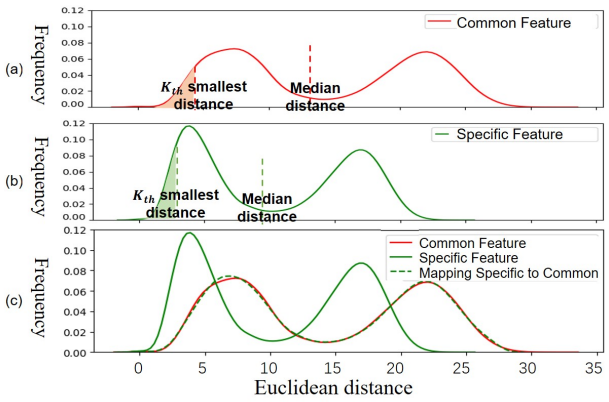

Figure 3: Distance distributions of (a) common feature, (b) specific feature, and (c) mapping specific feature to common feature. The mapped common feature distance distribution well fits the real common feature distance distribution.

metrics in two feature spaces are different. The distances can only be compared under the same evaluation metric. Thus we design an effective feature metric alignment scheme f_{align} before ranking, as follows.

Feature Distance Metric Mapping. The feature distance metric can be obtained from distance distribution. If two distributions are the same, consistent metric values can be obtained under any metrics. Therefore, we first analyze the feature distribution statistics in training set. We sample a large number of image pairs and calculate pair distances in two feature spaces. As shown in Fig. 3(a)(b), The median of the distances for common features is 13.67, while for the specific features is 9.85, *i.e.*, half of the distances are less than 13.67 and 9.85 in two feature space respectively. Therefore, the distance 13.67 for common features and 9.85 for specific features are the same evaluation metric scale in two feature spaces. The median is the special case of k_{th} smallest. More generally, the k_{th} smallest distance values can be selected to calculate mapping dictionary. By dense sampling, we obtain the mapping dictionary (key-value pairs) between common and specific feature distances.

Chebyshev Polynomials. The obtained mapping dictionary is discrete, but we hope to obtain a continuous mapping function to fit it. Thus we take a further step to fit it by using Chebyshev polynomials[Rivlin, 1974] that can provide an effective near-optimal approximation under the maximum norm. The Chebyshev polynomials of the first kind are defined by the recurrence relation:

$$\begin{aligned} T_0(x) &= 1, \qquad T_1(x) = x, \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x). \end{aligned} \tag{6}$$

Given the mapping dictionary, the Chebyshev approximation is to find the coefficients $\{c_0, ..., c_n\}$ to represent the mapping relation,

$$f(\theta) \sim \sum_{i=0}^{\infty} c_i T_i(\theta). \tag{7}$$

By truncating the function with n terms, we have the approximation function $\tilde{f}(\theta) = \sum_{i=0}^{n} c_i T_i(\theta)$ to fit the map-

VeRI-776				VehicleID					
Settings	Test Size= 11579			Settings	Query Number= 800			Query Number= 2400		
Methods	mAP	r = 1	r = 5	Methods	mAP	r = 1	r = 5	mAP	r = 1	r = 5
FACT +Plate + STR [Liu *et al.*, 2016c]	27.77	61.44	78.78	Mixed Diff [Liu *et al.*, 2016a]	54.6	48.93	75.65	45.5	41.05	63.38
VAMI [Zhou and Shao, 2018]	50.13	77.03	90.82	VAMI [Zhou and Shao, 2018]	-	63.12	83.25	-	47.34	70.29
EALN(VCCM) [Lou *et al.*, 2019a]	57.44	84.39	94.05	Defense Triplet[Hermans *et al.*, 2017]	68.9	65.2	77.93	61.37	57.20	71.91
FDA-Net (VGGM) [Lou *et al.*, 2019b]	55.49	84.27	92.43	FDA-Net (VGGM) [Lou *et al.*, 2019b]	68.94	65.91	86.15	61.84	55.53	74.65
RNN-HA(Resnet50) [Wei *et al.*, 2018]	56.80	74.79	80.51	RNN-HA(Resnet50) [Wei *et al.*, 2018]	-	68.8	81.9	-	62.6	77.0
AAVER(Resnet50) [Pirazh *et al.*, 2019]	58.52	88.68	94.10	AAVER(Resnet50) [Pirazh *et al.*, 2019]	-	70.03	89.81	-	59.04	80.60
VANet(Googlenet) [Chu *et al.*, 2019]	66.34	89.78	95.99	HDC + Contrastive [Yuan *et al.*, 2016]	65.5	-	-	57.5	-	-
MLSL(Mobilenet) [Alfasly *et al.*, 2019]	61.13	90.04	96.00	MLSL(Mobilenet) [Alfasly *et al.*, 2019]	-	74.21	88.38	-	66.55	78.67
PAMTRI(Dense201) [Tang *et al.*, 2019]	71.88	92.86	96.97	EALN(Resnet50) [Lou *et al.*, 2019a]	77.5	75.11	88.09	71.0	69.30	81.42
Specific feature only	67.72	89.39	95.94	Specific feature only	76.85	69.04	93.23	69.69	61.42	88.55
Common feature only	70.28	91.06	96.54	Common feature only	79.18	76.28	92.95	73.62	66.81	88.54
Simple combination of two features	71.39	91.49	97.13	Simple combination of two features	79.62	75.71	91.30	73.94	66.36	89.29
DFLNet (Resnet50)	**73.29**	**93.21**	**97.56**	DFLNet (Resnet50)	**82.83**	**78.83**	**95.01**	**75.40**	**69.78**	**90.59**

Table 1: Performance comparisons (%) with state-of-the-art methods on VeRI-776 and VehicleID datasets.

ping dictionary. In our work, when the order n is set to 4, the mapping can be well fitted.

The mapped common feature distance distribution is illustrated in Fig. 3(c), which well fits the real common distance distribution. The KLD (Kullback-Leibler divergence) calculated by these two distributions is 0.007. It shows the proposed feature metric alignment is reasonable and precise.

3.4 Implementation Details

The backbone of DFLNet is ResNet50 [He *et al.*, 2016]. DFLNet has two branches after "pool5" layer. One branch is two fully-connected (*fc*) layers with $\mathcal{L}_{emb}$ in Eq. 4 for specific features. The attention module in specific branch consists of two 1×1 convolution layers (channel 1st layer: $2048 \rightarrow 512$, 2nd layer: $512 \rightarrow 1$). The other branch is an embedding layer with adversarial learning for common features. The dimensions of the common and specific features are both 128. We adopt a hard example mining mechanism to obtain a strong baseline. The discriminator for common feature learning is a small network with three *fc* layers followed by a classifier. The number of classifier output is the size of odd-one unit, and we set it to 3. The input channels for each *fc* layer are 384 (128x3), 128 and 128. The DFLNet is optimized by SGD algorithm. Regarding parameters, we set ω as 0.5 and triplet margin as 0.6 in metric learning following [Lou *et al.*, 2019b], and $\lambda = 0.5$ in hybrid ranking. The models are trained for 50 epochs. Learning rate starts from 0.003. The size of the input image is 256×256.

4 Experiments

4.1 Experiment Setting

Dataset. We conduct experiments on VehicleID [Liu *et al.*, 2016a], VeRI-776 [Liu *et al.*, 2016c] and VERI-Wild [Lou *et al.*, 2019b] datasets, which are widely used vehicle ReID benchmarks. **VehicleID** consists of 26,267 vehicle IDs, and most of the vehicles only have two views: front-view and rear-view. **VeRI-776** is a small-scale vehicle dataset containing 776 vehicle IDs, which are captured by 20 cameras in unconstrained traffic scenarios. **VERI-Wild** is a large-scale vehicle dataset, which contains in total 416,314 images of 40,671 IDs captured by 176 surveillance cameras in the wild. The images in VehicleID dataset have 2 (front, back) orientations, and images in VeRI-776 and VERI-Wild datasets contain 5 (front, front-side, side, back-side, back) orientations.

Evaluation Metrics. We use mean Average Precision (mAP) and Cumulative Match Curve (CMC) in experiments.

4.2 Experimental Results Analysis

Evaluation on VehicleID Dataset

The results on VehicleID dataset are shown in Table 1. Using specific features only, we can achieve a strong baseline with the attention module and the hard example mining scheme. Our DFLNet outperforms EALN method by 5.33% (mAP). EALN uses a hard negative generation scheme for discriminative feature learning and a cross-view generation scheme to improve cross-view vehicle ReID. In Top 5 recall, DFLNet can achieve 9% mAP advantages over EALN. Besides, HDC+Contrastive cascades a set of GoogleNet which is a more complex method than a single network in our model, which further demonstrates the effectiveness of DFLNet. VAMI aims to infer viewpoint aware attentive regions for multi-view feature representation and AAVER proposes an adaptive attention mechanism to capture discriminative features. Compared with these methods, DFLNet achieves better performance with a disentangled learning scheme.

Evaluation on VeRI-776 Dataset

The results on VeRI-776 dataset are shown in Table 1. Compared to using specific features, the proposed DFLNet can significantly improve the performance from 67.72% to 73.29% mAP via introducing another common feature. Such gains demonstrate that by disentangling specific and common features, more adaptive feature representation and matching can be achieved. Our DFLNet significantly outperforms the state-of-the-art method PAMTRI, and it is worthy noting that PAMTRI uses DenseNet201 and the dimension of feature is 1024, while we use a simple ResNet50 network and two 128-dimension features in DFLNet. Such observation verifies that DFLNet can better promote feature representation.

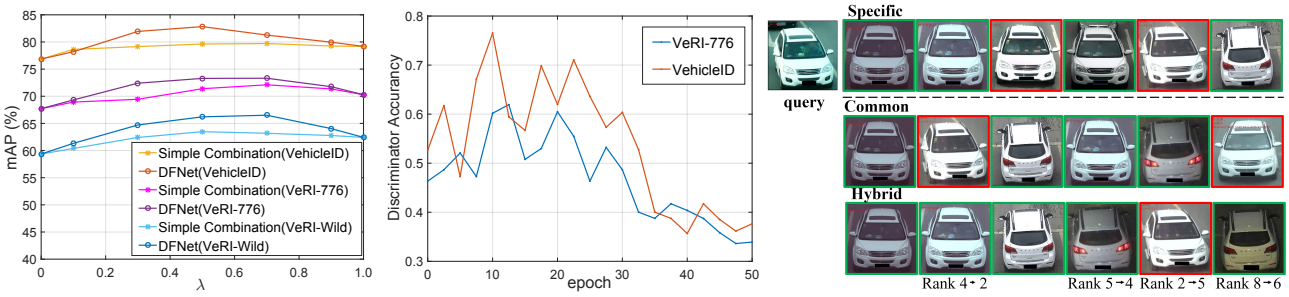

Figure 4: Left(a): the ReID performance by setting different scale λ in the ranking strategy. Middle(b): the discriminator classification accuracy in "odd-one-out" adversarial learning. Right(c): the Top 6 ReID results on VehicleID dataset. The green/red boxes indicate right/wrong results. Compared with common feature results, the rank changes after using our hybrid ranking scheme.

Settings	Small	Medium	Large
FDA-Net (VGGM)[Lou *et al.*, 2019b]	35.11	29.80	22.78
Softmax (Resnet50) [Liu *et al.*, 2016c]	49.76	41.28	30.91
Triplet (Resnet50) [Schroff *et al.*, 2015a]	57.69	46.81	34.73
MLSL (Mobilenet) [Alfasly *et al.*, 2019]	46.32	42.37	36.61
FDA-Net (Resnet50) [Lou *et al.*, 2019b]	61.57	52.69	45.78
Specific feature only	59.36	50.81	39.61
Common feature only	62.42	51.84	42.97
Simple combination of two features	63.48	52.91	43.05
Hybrid ranking w/o alignment	50.12	40.86	27.89
DFLNet	**66.21**	**58.28**	**47.16**

Table 2: The mAP performance on the VERI-Wild dataset.

Evaluation on VERI-Wild Dataset

The results on VERI-Wild dataset are illustrated in Table 2. Compared with FDA-Net that uses GAN to generate hard negative samples, we get better performance under same network backbone. Moreover, we show the comparisons between hybrid ranking and simple combination of common and specific features. Simple combination directly weighted two distances of common and specific features, *i.e.* $\lambda d_c + (1-\lambda)d_s$. As shown in Tables 1 and 2, the simple combination can also bring performance gains, but is lower than hybrid ranking. We vary the value of λ in ranking scheme as shown in Fig. 4(a). The hybrid ranking can get consistent performance superiority over different λ. Since specific features focus on subtle differences, using it for cross-view ReID may degrade the cross-view ReID performance. We also provide the results of hybrid ranking without alignment, the performance drops significantly. This is mainly due to the feature distance distribution differences between these two features.

Ablation Study of Common and Specific Features

In Table 3, we present the ablation results of DFLNet. EALN [Lou *et al.*, 2019a] is the first work that reports the same-view and cross-view performances. Specifically, given a query vehicle, for cross-view ReID, we treat the reference vehicles belonging to this query but with the same view as the junk samples, which are not involved in mAP computation. Simply comparing specific and common features, it can be observed that specific features are good at finding the same view vehicles while the common features are good at finding cross-view vehicles. Such results show the effectiveness of our disentangled feature learning scheme.

In Fig. 4(c), we visualize the retrieval results. During testing, we use common features to get initial recall list (second row), then use specific features to further compare samples with same orientation as query to get hybrid ranking results (third row). The wrong samples with the same orientation as query can be further filtered by specific features. Therefore, the performance of hybrid ranking is superior than only using specific or common features in both same and cross view.

Methods	ALL	Same View	Cross View
EALN[Lou *et al.*, 2019a]	77.5	89.2	42.6
Specific feature	76.85	87.19	46.67
Common feature	79.18	84.29	56.83
DFLNet	**82.83**	**90.22**	**58.69**

Table 3: The ReID performance (mAP) of same-view and cross-view on the small scale (TestSize=800) test set in VehicleID.

The "Odd-One-Out" Adversarial Learning Analysis

We design a toy experiment to verify whether orientation specific features learned by embedding network contain enough orientation specific information. For VehicleID and VeRI-776 dataset, the "Odd-One-Out" classification accuracy is up to 93.28% and 87.69%, respectively. That means for well-trained embedding networks, the specific information is implicitly encoded in the features and the discriminator has the ability to recognize the odd one in given specific supervision.

For adversarial learning, as the training continues, the classification accuracy for odd-one prediction gradually decreases, as shown in Fig. 4(b). Thus, the expected orientation common features can be learned.

5 Conclusion

In this paper, we focus on exploiting the common and specific features to improve vehicle ReID performance. A novel disentangled feature learning network is proposed to jointly learn these two features. Such a unified end-to-end solution lays the groundwork for the subsequent improvements of vehicle ReID in terms of discriminative and invariant feature learning. Experiments demonstrate the effectiveness of the proposed method.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant U1611461, and was partly supported by the SUTD SRG project (T1SRIS20153).

References

[Alfasly *et al.*, 2019] Saghir Alfasly, Yongjian Hu, Haoliang Li, Tiancai Liang, Xiaofeng Jin, Beibei Liu, and Qingli Zhao. Multi-label-based similarity learning for vehicle re-identification. *IEEE Access*, 7:162605–162616, 2019.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and et al. Neural machine translation by jointly learning to align and translate. *Computer Science*, 2014.

[Bai *et al.*, 2018] Yan Bai, Yihang Lou, and et al. Group sensitive triplet embedding for vehicle re-identification. *IEEE Transactions on Multimedia*, 2018.

[Bulan *et al.*, 2017] Orhan Bulan, Vladimir Kozitsky, Palghat Ramesh, and Matthew Shreve. Segmentation-and annotation-free license plate recognition with deep localization and failure identification. *IEEE Transactions on Intelligent Transportation Systems*, 18(9):2351–2363, 2017.

[Chu *et al.*, 2019] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, and et al. Vehicle re-identification with viewpoint-aware metric learning. In *IEEE International Conference on Computer Vision*, pages 8282–8291, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2019] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019.

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

[Liu *et al.*, 2016a] Hongye Liu, Yonghong Tian, and et al. Deep relative distance learning: Tell the difference between similar vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.

[Liu *et al.*, 2016b] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *IEEE International Conference on Multimedia and Expo, 2016*, pages 1–6. IEEE, 2016.

[Liu *et al.*, 2016c] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *IEEE International Conference on European Conference on Computer Vision*, pages 869–884. Springer, 2016.

[Lou *et al.*, 2019a] Yihang Lou, Yan Bai, and et al. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 2019.

[Lou *et al.*, 2019b] Yihang Lou, Yan Bai, and et al. Veriwild: A large dataset and a new method for vehicle re-identification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019.

[Ma *et al.*, 2017] Liqian Ma, Qianru Sun, and et al. Disentangled person image generation. *arXiv preprint arXiv:1712.02621*, 2017.

[Pirazh *et al.*, 2019] Khorramshahi Pirazh, Amit Kumar, Neehar Peri, and et al. A dual path modelwith adaptive attention for vehicle re-identification. *IEEE International Conference on Computer Vision*, 2019.

[Rivlin, 1974] Theodore J Rivlin. *The Chebyshev polynomials*. Wiley, 1974.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, and et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[Schroff *et al.*, 2015a] Florian Schroff, Kalenichenko Dmitry, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[Schroff *et al.*, 2015b] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, and et al. Rethinking the inception architecture for computer vision. In *the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Tang *et al.*, 2019] Zheng Tang, Milind Naphade, and et al. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *IEEE International Conference on Computer Vision*, pages 211–220, 2019.

[Tran *et al.*, 2017] Luan Tran, Xi Yin, and et al. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017.

[Wang *et al.*, 2017] Zhongdao Wang, Luming Tang, and et al. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *IEEE International Conference on Computer Vision*, pages 379–387, 2017.

[Wei *et al.*, 2018] Xiu-Shen Wei, Chen-Lin Zhang, and et al. Coarse-to-fine: A rnn-based hierarchical attention model for vehicle re-identification. In *Asian Conference on Computer Vision*, pages 575–591. Springer, 2018.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, and et al. Show, attend and tell: Neural image caption generation with visual attention. *Computer Science*, pages 2048–2057, 2015.

[Yuan *et al.*, 2016] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. *arXiv preprint arXiv:1611.05720*, 2016.

[Zhao *et al.*, 2019] Yiru Zhao, Xu Shen, and et al. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[Zhou and Shao, 2018] Yi Zhou and Ling Shao. Viewpoint aware attentive multi-view inference for vehicle re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2018.