Unstructured Feature Decoupling for Vehicle Re-Identification

Wen Qian[1,2], Hao Luo[3], Silong Peng[1,2], Fan Wang[3], Chen Chen[1]*, and Hao Li[3]

[1] Institute of Automation, Chinese Academy of Sciences
{qianwen2018,silong.peng,chen.chen}@ia.ac.cn
[2] University of Chinese Academy of Sciences, School of Artificial Intelligence
[3] Alibaba group {michuan.lh,fan.w,lihao.lh}@alibaba-inc.com

Abstract. The misalignment of features caused by pose and viewpoint variances is a crucial problem in Vehicle Re-Identification (ReID). Previous methods align the features by structuring the vehicles from predefined vehicle parts (such as logos, windows, etc.) or attributes, which are inefficient because of additional manual annotation. To align the features without requirements of additional annotation, this paper proposes a **Unstructured Feature Decoupling Network** (UFDN), which consists of a transformer-based feature decomposing head (TDH) and a novel cluster-based decoupling constraint (CDC). Different from the structured knowledge used in previous decoupling methods, we aim to achieve more flexible unstructured decoupled features with diverse discriminative information as shown in Fig. 1. The self-attention mechanism in the decomposing head helps the model preliminarily learn the discriminative decomposed features in a global scope. To further learn diverse but aligned decoupled features, we introduce a cluster-based decoupling constraint consisting of a diversity constraint and an alignment constraint. Furthermore, we improve the alignment constraint into a modulated one to eliminate the negative impact of the outlier features that cannot align the clusters in semantics. Extensive experiments show the proposed UFDN achieves state-of-the-art performance on three popular Vehicle ReID benchmarks with both CNN and Transformer backbones. Our code is released at: https://github.com/damo-cv/UFDN-Reid.

Keywords: Unstructured Feature Decoupling Network, Vehicle ReID, Transformer-based Decoupling Head, Cluster-based Decoupling Constraint.

1 Introduction

Given a query vehicle image, Vehicle ReID aims to retrieve images of the same vehicle from the gallery that contains images captured by disjoint cameras. With the development of large Vehicle ReID benchmarks [18,20,13,40,14,31] and deep

[*] Corresponding author is Chen Chen.
The work was supervised by Hao Luo and Chen Chen.

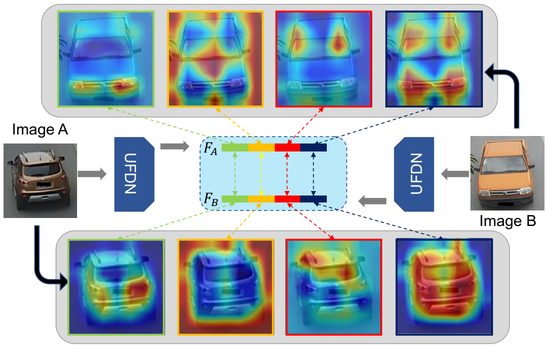

Fig. 1. Illustration of the decoupling and alignment process for vehicle features.

learning methods, Vehicle ReID achieves a great promotion in performance and is widely applied in intelligence city system [20,5,27]. However, it is challenging to deal with the misalignment of features caused by poses.

Existing methods [32,1,4,43,3,17] mainly tackle the misalignment of features by structuring the vehicles in two directions. 1) Some methods [32,25,36,26] spatially decompose the vehicle into several stripes or grids. Specifically, PCB [32] and its variants divide the feature maps into multi-level stripes/grids straightforwardly to integrate fine-grained information. However, these methods only decompose the feature from the spatial dimension, which is coarse and lacks semantic representations. 2) Recent methods [1,4,43,3,17] utilize prior semantic information such as pre-defined vehicle parts (e.g. lights, windows, etc.) and vehicle attributes (colors, viewpoints, etc.) to guide the feature decoupling. For example, PVEN [24] decouples the vehicle features in the view-aware feature space based on a parsing network and the pre-defined viewpoint labels. Nevertheless, such explicit alignment based on the pre-defined knowledge cannot flexibly handle missing components caused by viewpoint variances or occlusions. Structured analysis of vehicle images is usually time-consuming and inefficient since they rely heavily on manual annotation and extra modules [8,19,28]. Moreover, the local feature explored in the above methods will be only discriminative for the corresponding region due to the hand-crafted rules.

To address the aforementioned limitations, this paper studies the implicit alignment of features by first decomposing them into unstructured parts and then aligning them without using extra annotation. However, there mainly exist two challenges: 1) how to decompose the feature without using extra structured cues; 2) how to learn diverse but aligned decoupled features?

We propose a transformer-based decomposing head (TDH) to decompose the vehicle feature into unstructured parts. Different from those methods [32,25,36,26] that factorize the images on the spatial dimension, TDH keeps a global receptive field of each decomposed feature through decomposing the feature map from the channel dimension. We feed each decomposed group of feature maps into a modified transformer block, and then the self-attention mechanism can automatically encode one discriminative feature in a global scope. Since the feature map is not simply divided into fixed stripes/grids, the decomposed features can implicitly learn discriminative semantic information without extra cues.

Apart from the implicit decomposing module, we propose a novel cluster-based decouple constraint (CDC) to improve the diversity and alignment of decomposed features in an annotation-free way. CDC aims to cluster the decomposed features into groups, which consists: 1) the diversity constraint: the decomposed features should be orthogonal to each other, which motivates them to focus on diverse regions of interest; 2) the alignment constraint: the decomposed features should be close to the relevant cluster centers to align with each other. However, some outlier features cannot align with the cluster centers in semantics, which will lead to useless or even inaccurate supervision. To tackle such an issue, we filter out the outlier features to mitigate their negative effects and term the final output as the decoupled features. Different from that methods [16,2] such as ABDNet (see the details in Section 5), which only conduct the diversity constraint on the 2D features maps to keep the diversity of features, our UFDN aims to maintain the diversity and alignment of final 1D features.

Moreover, we visualize samples from different viewpoints as shown in Fig. 1 and find that the corresponding decoupled features tend to focus on similar salient regions (e.g., the lights information in the first part, the counter, and front information for the second part, etc.) We term our method as unstructured feature decoupling network (UFDN), and experiment on three popular benchmarks with two different backbones (ResNet and Swin-transformer) to evaluate the effectiveness of UFDN. **The contributions of this paper are:**

i) We propose UFDN which aims to alleviate the feature misalignment in Vehicle ReID by decoupling them into unstructured, diverse, and aligned parts without human annotation.
ii) The transformer-based unstructured feature decomposing head can decompose features into several groups from the channel dimension in a global scope which is more robust than the local specified methods.
iii) We propose a cluster-based decoupling constraint to keep the diversity and alignment of the decoupled features without human annotation and further external the outliers to eliminate the negative influence on them.
iv) Without the requirement of extra manual annotation, our UFDN outperforms other methods on three popular benchmarks consistently.

2 RELATED WORK

Previous methods [5,18,1,4,43,7,11,17] decouple vehicle features for better alignment and can be categorized into two kinds:

The Spatial Decomposing Methods [32,25,36,26]. These methods spatially decompose the feature map into several stripes or grids to capture fine-grained information. Sun et al. [32] propose to decompose the feature map into six parts and each local feature is followed by the ReID supervision. Mo et al. [25]

employ a cascaded hierarchical context-Aware Vehicle ReID network to decompose the vehicle features at multi-scale hierarchically. The coarse decomposed features provided by the spatial decomposing methods have achieved progress in ReID performance. However, the improvement brought by the spatial-wise decomposing is limited since it tends to focus on coarse local regions and neglect the important semantic information.

The Pre-defined Semantic Decoupling Methods Different from the coarse decomposed information in previous spatial decomposing methods, the predefined semantic decoupling methods [1,4,43,7,11,17,24,6,39] decouple the vehicle feature with the assistance of the pre-defined information, e.g., pre-defined vehicle parts (wheels, logos, and windows) or vehicle attributes (color, type, and viewpoints). Zhang et al. [37] introduce a part-guided attention network to enhance the feature representation by decoupling the features under the guidance of the prominent parts. Apart from the decoupling methods guided by vehicle parts, Wang et al. [35] propose an attribute-guided module to assist the decoupling of features. Guo et al. [6] bridge the gap between the vehicle features from different models by a coarse-to-fine structured feature embedding. However, the above methods [1,17,24,25,36,26] rely heavily on human annotation and only focus on the fixed pre-defined regions while ignoring other potential crucial clues.

Most existing methods decouple features in a structured way, we target to study an annotation-free method that decouples the vehicle features into unstructured, diverse, and aligned ones.

3 METHODOLOGY

Fig. 2 shows the illustration of UFDN, which mainly consists of a transformer-based feature decomposing head and a cluster-based decoupling constraint.

3.1 Backbone and Symbol Definition

Given an input image X, the backbone outputs a feature map which is reshaped to a base feature $F_{base} \in R^{n \times c}$, where $n = H \times W$ and c represent the spatial dimension and the channel dimension, respectively. Then we split F_{base} into k groups from the channel dimension to obtain a new feature set $F_p \in R^{k \times n \times m}$, where $c = k \times m$. The backbone can be both CNN-based or Transformer-based.

3.2 Transformer-based Feature Decomposing Head

The transformer-based feature decomposing head (TDH) encode the unstructured information of each decomposed feature $F_p^i \in R^{n \times m}, i = 1, 2, 3, ..., k$. As shown in Fig. 2, k decomposing tokens $T^i \in R^m$ are pre-pended to the relative channel-wise feature $F_p^i \in R^{n \times m}$, respectively. The input sequence fed to transformer-based feature decomposing head (TDH) is denoted as $z_0^i = [T^i, F_p^i]$.

As illustrated in Fig. 3, TDH contains a total of L transformer blocks each of which consists of a multi-head decomposing attention (DA) module and a

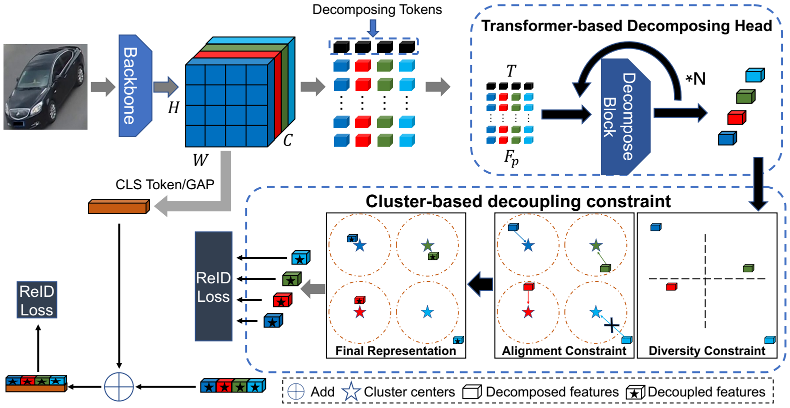

Fig. 2. Illustrative the framework of UFDN consisting of two modules: 1) the transformer-based feature decomposing head that aims to learn unstructured decomposed features of the original base feature from the attention mechanism; 2) the cluster-based decouple constraint that aims to keep the diversity and alignment of the decoupled features. Moreover, the base-feature extraction module can be either a CNN-based network (ResNet) or a Transformer-based network (Swin transformer).

MLP module. Since F_p^i is a deep-layer feature that is good enough for encoding discriminative information, we follow the solution [34] to update only the decomposing token T^i to re-aggregate F_p^i, *i.e.* F_p^i is frozen during training to reduce computational cost. Given the input sequence $z_0^i = [T^i, F_p^i]$, we term the output decomposing token of the $l-1$ block in TDH as T_{l-1}^i and the input sequence to the l-th block as $z_{l-1}^i = [T_{l-1}^i, F_p^i]$. Then we will feed the l-th input sequence z_{l-1}^i to the DA module of the l-th block in TDH, which can be expressed as:

$$\begin{aligned} &Q = W_q T_{l-1}^i,\ K = W_k z_{l-1}^i,\ V = W_v z_{l-1}^i, \\ &A = Softmax(QK^T),\ h_l^i = A \cdot V + T_{l-1}^i, \end{aligned} \tag{1}$$

where $W_q, W_k, W_v \in R^{m \times m}$ are the projection matrices. After getting hidden variable h_l^i from the multi-head decomposing attention module, we feed it to the MLP module in the l-th TDH block as follows:

$$T_l^i = LN(MLP(LN(h_l^i)) + h_l^i), \tag{2}$$

where MLP and LN denote the MLP module and the layer normalization layer, respectively. Then, the decomposing token T_l^i will be concatenated with F_p^i as the input sequence z_{l+1}^i for the next block.

We integrate the decomposing token $T_o^i = T_L^i$ outputted by the last block in TDH and the relevant global feature as the decomposed feature $F^i \in R^m$:

$$F^i = T_o^i + GAP(F_p^i), \tag{3}$$

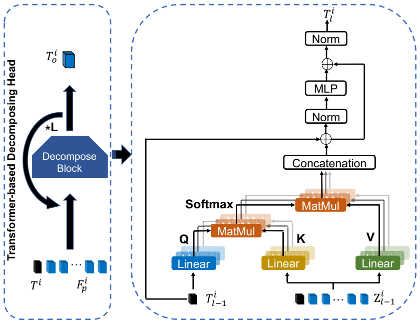

Fig. 3. The transformer-based feature decomposing head: the divided feature F_p^i and decomposing token T^i are fed into the block to calculate the composed feature T_o^i.

where GAP represents the global average pooling. Finally, k decomposed features $[F^1, F^2, ..., F^k]$ are concatenated to the final ReID feature $F \in R^c$. We conduct the ReID loss on each decomposed feature and the ReID feature as follows:

$$\begin{aligned} L_{THD} &= L_{ReID}(F) + \sum_{i=1}^{k} L_{ReID}(F^i), \\ L_{ReID} &= L_{ce} + L_{tri}, \end{aligned} \tag{4}$$

where L_{ce} represents the cross entropy loss and L_{tri} represents the triplet loss.

3.3 The Cluster-based Decoupling Constraint

Apart from the implicit feature decomposing module, a cluster-based decouple constraint (CDC) is proposed to obtain the decoupled features by requiring the diversity and alignment of these decomposed features as shown in Fig. 2. Firstly, we employ a diversity constraint to enforce the diversity of the decomposed features. Secondly, to align the decomposed features between different images, we propose an alignment constraint that clusters the relevant decomposed features into groups and eliminates the negative effects of outlier features.

The Diversity Constraint. Given an input image X, we obtain k decomposed features $F = [F^1, F^2, ..., F^k]$ and hope that each of them should have diverse regions of interest (ROIs) semantically. Enforcing the diversity of decomposed features can drive the model to mine more salient and discriminate information and accelerate the decoupling process of vehicle features. For abstracting this requirement into a mathematical description, we constrain these features to be orthogonal to each other. The diversity constraint restricts the Gram matrix of F to be close to an identity matrix under Frobenius norm:

$$L_{div} = ||FF^T - I||_F \tag{5}$$

The Alignment Constraint. The diversity constraint only focuses on the relationship between features from the same image, which neglects the cross-image relationship. So we propose an alignment constraint for ranking the decomposed features from different images in the same order. For example, given two input images X_1 and X_2, we obtain the corresponding decomposed features $F_1 = [F_1^1, F_1^2, ..., F_1^k]$ and $F_2 = [F_2^1, F_2^2, ..., F_2^k]$, and hope both the blocks F_1^1 and F_2^1 can focus on the same vehicle region. We propose to build a cluster center for each group of decomposed features and require the decomposed features should be close to the relevant cluster centers to align with each other.

Clustering of Decomposed Features. Given M samples of the training set, we decompose each of them into k decomposed features and build a cluster center C^i for each group of decomposed features $[F_1^i, F_2^i, ..., F_M^i],\ i \in [1, k]$:

$$C^i = \frac{1}{M}\sum_{j=1}^{M} F_j^i, \tag{6}$$

where F_j^i is the i-th decoupled feature of the image X_j. During the early period of the training process, the decomposed features experience severe fluctuations caused by the large network adjustment from the loss backpropagation, which leads to an unstable convergence process of the cluster centers. To smooth the convergence of cluster centers, a memory bank is adopted to store the cluster centers updated with a momentum strategy.

$$C_t^i = \alpha C_{t-1}^i + (1-\alpha)\frac{1}{M}\sum_{j=1}^{M} F_j^i, \tag{7}$$

where α is weight for controlling the update speed of the cluster centers, and t denotes different periods of the training process. We treat the cluster-center C^i as a standard feature center since it has walked through all samples for epochs and tends to learn a more general and comprehensive feature representation.

After building the cluster centers for each group of features, we constrain the distance between the positive pair (F_j^i and its corresponding cluster center C^i) should be smaller than the smallest distance of negative pairs (other cluster centers). It is termed as the alignment constraint:

$$L_{align} = [d(F_j^i, C^i) - \min(d(F_j^i, C^{q\in[1,k]),q\neq i})) + \theta]^+, \tag{8}$$

where $d(a, b)$ measures the distance between a and b, $[\cdot]^+ = max(\cdot, 0)$, and θ is the distance threshold to control the distance between the positive and negative pairs. After constraining the decomposed features by the diversity constraint and alignment constraint, we term them as decoupled features.

Exclusion of Outlier Features. Some decoupled features may lack the semantic information of the cluster-center due to different viewpoints or poses, and we term them as outlier features. The outlier features cannot align the cluster centers in semantics, e.g., the front window information is absent in the images captured from the backside, which will result in performance degradation.

So it is necessary to eliminate the negative impact from the outlier features by excluding the corresponding alignment loss back-propagation of them.

Firstly, we calculate the average distance D_i of each center C^i:

$$D_i = \frac{1}{M}\sum_{j=1}^{M} d(F_j^i, C^i). \tag{9}$$

Then we will exclude the loss from the feature block F_j^i if the distance between it and the corresponding cluster center is greater than the average distance D_i. After excluding the outlier features, the modulated alignment loss is:

$$L_{mod} = \begin{cases} L_{align}, & if\ d(F_j^i, C^i) < D_i \\ 0, & if\ d(F_j^i, C^i) \geq D_i. \end{cases} \tag{10}$$

The cluster-based decouple constraint is computed as:

$$L_{CDC} = \frac{1}{M}\sum_{j=1}^{M}[L_{div}(F_j) + \sum_{i=1}^{k} L_{mod}(F_j^i, C^i, D_i)]. \tag{11}$$

Finally, the loss of the UFDN is: $L = L_{THD} + L_{CDC}$

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Dataset. We experiment on three Vehicle ReID benchmarks: VeRi776 [18], VehicleID [20] and VERI-WILD [22]. VeRi776 [18] is a classic Vehicle ReID benchmark and contains 776 identities collected by 20 cameras in a real-world environment. VehicleID [20] is a large-scale dataset collected by multiple cameras during the daytime on the open road, which contains 26,267 vehicles and 221,763 images in total. VERI-WILD [22] is another large-scale dataset, and it consists of 40,671 vehicles and 416,314 images. Moreover, VERI-WILD [22] is collected by 174 cameras during a month which is a long period.

Evaluation Metrics. We use the same evaluation protocols used in previous methods [20,41,38] for evaluation: Mean Average Precision (mAP) and the cumulative matching characteristics at Rank1 (CMC@1). Moreover, VehicleID [20] reports only CMC@1 because mAP is unavailable due to its unique test sets.

4.2 Implementation Details

We perform experiments in PyTorch on a machine with 8 NVIDIA V100 GPU. The images are resized to 224 × 224 for both training and testing, and the augmentation includes random erasing and flipping. We train the modules of UFDN together with a warmup strategy [9] and adapt different optimizers for different backbones: 1) Adam optimizer with the weight decay factor of 1e-4 for

Table 1. Comparison with state-of-the-art methods. It includes mAP and CMC@1 on VeRi-776; CMC@1 and mAP on three test sets of small (S), medium (M), and large (L) on VehicleID and VERI-WILD, respectively. Baseline* and UFDN* are CNN-based. Baseline and UFDN are Transformer-based. Finally, "/" indicates numbers were not reported. For a fair comparison, all methods report the results of the same-scale backbones (*i.e.* ResNet50, ViT-Small, and Swin-Tiny) pre-trained on ImageNet-1K.

		VeRi-776		VehicleID			VERI-WILD		
Method	Backbone	mAP	CMC@1	CMC@1 (S)	CMC@1 (M)	CMC@1 (L)	mAP (S)	mAP (M)	mAP (L)
PGAN [37]	Res50	79.3	96.5	77.8	/	/	74.1	/	/
PRN [7]	Res50	74.3	94.3	78.4	75.0	74.2	/	/	/
PVEN [24]	Res50	79.5	95.6	84.7	80.6	77.8	82.5	77.0	69.7
GLAMOR [33]	Res50	80.3	96.5	78.6	/	/	77.2	/	/
AGNet-ASL [35]	Res50	71.6	95.6	71.2	69.2	65.7	/	/	/
SAN [26]	Res50	72.5	93.3	79.7	78.4	75.6	/	/	/
AAVER [12]	Res50	61.2	89.0	74.7	68.6	63.5	/	/	/
SEVER [13]	Res50	79.6	96.4	79.9	77.6	75.3	83.4	78.7	71.3
VAMI [42]	Res50	61.3	89.5	63.1	52.9	47.3	/	/	/
DCDLearn [44]	Res50	70.4	92.8	82.9	78.7	75.9	/	/	/
CAL [29]	Res50	74.3	95.4	82.5	/	/	/	/	/
GB+GFB+SLB [15]	Res50	81.0	96.7	86.8	/	/	/	/	/
TransReID [10]	Vit-base	78.0	96.1	82.9	/	/	/	/	/
TransReID [10]	Swin-tiny	77.2	95.6	80.5	/	/	/	/	/
Baseline	Swin-Tiny	78.2	95.6	84.6	80.9	77.5	80.7	76.3	69.1
Baseline*	Res50	79.6	95.6	85.7	82.2	78.8	81.8	77.1	69.9
UFDN	Swin-Tiny	80.9	96.3	85.9	82.4	79.3	82.0	77.5	70.4
UFDN*	Res50	**81.5**	96.4	**88.4**	**84.8**	**80.6**	**84.6**	**79.4**	**72.0**

UFDN with the CNN backbone; 2) AdamW optimizer [21] with the weight decay factor of 1e-4 for UFDN with the Transformer backbone. For both optimizers, we initialize the learning rate as 3e-4. The hyper-parameter α in the memory bank mechanism is set to 0.9. Moreover, we set the hyper-parameter θ as 0.1 to control the distance between the positive distance and the negative distance.

4.3 Comparisons to State-of-the-Art Methods

We compare our UFDN with a wide range of state-of-the-art methods as shown in Table 1, including (1) part-based approaches: PGAN [37], PRN [7], PVEN [24], and GLAMOR [33]; (2) attribute-based approaches: AGNet-ASL [35] and SAN [26]; (3) attention-based approaches: AAVER [12] and SEVER [13]; (4) other interesting approaches: VAMI [42], DCDLearn [44], GB+GFB+SLB [15], CAL [29] and TransReID [10], and we get the following conclusions:

1) The CNN-based UFDN* has already achieved state-of-the-art performance on all three benchmarks by aligning the unstructured decoupled vehicle features. The CNN-based baseline* reaches a 79.6% mAP on the VeRi-776 benchmark, and the UFDN* achieves an 81.5% mAP. Moreover, UFDN also achieves the best performance on VehicleID benchmarks that outperforms the second-best competitor GB+GFB+SLB [15] by 1.6% CMC@1 even though it borrows self-supervised representation learning to facilitate geometric features discovery.

2) We also compare UFDN with TransReID [10] (Swin-tiny). The baseline achieves 78.2% on VeRi-776 and the UFDN achieves a 2.7% mAP improvement.

Table 2. Ablation study of the components in UFDN (mAP on VeRi-776), where TDH represents the feature decomposing head and CDC represents the cluster-based decoupling constraint.

TDH	CDC	UFDN (Res50)	UFDN (Swin-Tiny)
×	×	79.6	78.2
✓	×	81.0	80.0
✓	✓	81.5	80.9

Table 3. Ablation study of the depth of THD, where we report mAP for VeRi-776 and CMC@1 for VehicleID.

	UFDN (Res50)		UFDN (Swin-Tiny)	
Depth	VeRi-776	VehicleID	VeRi-776	VehicleID
1	80.2	86.5	79.2	84.3
2	**81.5**	**88.4**	**80.9**	**85.9**
3	80.9	87.5	79.7	84.9
4	79.1	85.2	78.2	83.0

When comparing UFDN with the TransReID [10], we achieve a 3.7% improvement of mAP on VeRi-776 benchmark and a 5.4% improvement of CMC@1 on VehicleID benchmark. Please note that TransReID employs the camera and viewpoint labels as prior knowledge while our UFDN is annotation-free.

4.4 Ablation Study and Evaluation

Ablation Study on Components in UFDN. To evaluate the effectiveness of the two proposed modules in UFDN, we perform an ablation study by adding the components step-by-step in Table 2 on the VeRi-776 benchmark with either CNN-based or Transformer-based backbone. We keep all the hyper-parameters same to ensure a fair comparison, and get the conclusions:

1) After adding the TDH on the baseline, we observe the mAP on VeRi-776 increased by 1.4% and 1.8% on ResNet50 and Swin-Tiny backbones, respectively. It indicates that just decomposing the features from the channel dimension and then enhancing the feature representation by the transformer-based feature decomposing head can already bring a performance improvement.

2) Although TDH has already reached performance progress, it can't guarantee the diversity and alignment of the decomposed features which is realized by the CDC. We do ablation studies on CDC and find that the diversity and alignment constraint can drive an extra performance improvement over TDH, e.g., 0.5% and 0.9% mAP on the two backbones.

Ablation Study on Feature Decomposing Head. We use the self-attention mechanism in the feature decomposing head to preliminarily learn the decoupled features, and we do ablation studies on the depth of the decomposing head and the number of decoupled parts in this part.

1) The Depth of Feature Decomposing Head. In Table 3, we investigate the influence of the depth of the decomposing head on VeRi-776 and VehicleID benchmarks. Take UFDN (Res50) as an example: we achieve the best performance with the depth as 2, which achieves 88.4% CMC@1 on the VehicleID benchmark. But the models with a shallower depth (depth as 1) or a deeper one (depth as 4) show poor performance, which reaches 86.5% and 85.2% CMC@1 on the VehicleID benchmark, respectively. A similar phenomenon exists when

Table 4. Ablation study of the number of decoupled parts in THD, where D_{Num} represents the number of the decoupled features. Moreover, we report CMC@1 for VehicleID and mAP for VeRi-776.

	UFDN (Swin-Tiny)		UFDN (Res50)	
D_{Num}	VeRi-776	VehicleID	VeRi-776	VehicleID
1	79.5	84.1	80.0	86.5
2	80.2	85.2	80.5	87.3
4	**80.9**	**85.9**	**81.5**	**88.4**
8	80.1	85.3	80.6	88.1
16	79.6	85.0	80.3	87.8

Table 5. Ablation study of the components in CDC (mAP on VeRi-776), where DC denotes the diversity constraint, AC denotes the alignment constraint, and MAC denotes the modulated alignment constraint.

DC	AC	MAC	UFDN (Res50)	UFDN (Swin-Tiny)
×	×	×	81.0	80.0
✓	×	×	81.2	80.5
×	✓	×	80.8	79.8
×	×	✓	81.1	80.4
✓	✓	×	81.2	80.3
✓	×	✓	**81.5**	**80.9**

experimenting with UFDN (Swin-Tiny) on both VeRi-776 and VehicleID benchmarks. Reasons for the above experiments can be concluded as: 1) TDH with shallow depth mines insufficient information from the input features and thus provides a poor performance; 2) TDH with a deeper depth is hard to train since the transformer-based networks rely heavily on large scale pre-training and the TDH added in our work is not pre-trained.

2) The Number of Decoupled Parts. We set the dimension of the output feature as 2048 in all our experiments for a fair comparison, and thus the number of decoupled parts needs to be divisible by 2048. We do ablation studies on the number of the decoupled parts in THD as shown in Table 4, and achieve the best performance when decoupling the features into four parts. Just changing the number of decoupled parts, there can be a great performance improvement, e.g., the performance improved from 86.5% CMC@1 in 1-part decoupled to 88.4% CMC@1 in 4-part decoupled on the VehicleID benchmark. We achieve two conclusions: 1) the decoupled features are not discriminative enough if decoupling it into 2 parts, and thus it's difficult to align them. 2) the ReID loss on each decoupled feature is hard to converge when we decouple a vehicle feature into too many parts and each of them contains little useful information.

Ablation Study on Cluster-based Decoupling Constraint. We propose the cluster-based decoupling constraint to keep the diversity and alignment of the decoupled features, which consists of the diversity constraint, the alignment constraint, and the upgraded modulated alignment constraint. For figuring out how they impact the final performance, we do ablation studies on the three components and the hyperparameters in them.

1) Two steps in CDC. To explore the effect of the diversity constraint and the modulated alignment constraint, we do ablation studies as shown in Table 5: DC drives a consistent improvement on both UFDN (Res50) and UFDN (Swin-Tiny), but we should notice the different results after adding AC which shows a slight performance decrease on UFDN (Swin-Tiny). It is because the transformer-based network is more sensitive than the CNN-based network and can be easily affected by the negative impact from the outlier features.

Table 6. Ablation study of the alignment constraint, and $\theta = 0$ denotes soft margin.

			VeRi-776		VehicleID
Scheme	θ	α	mAP	CMC@1	CMC@1
Scheme 1	0	0.5	80.7	95.8	87.0
Scheme 2	0.1	0.5	80.9	96.4	87.5
Scheme 3	0.3	0.5	80.7	96.1	87.2
Scheme 4	0.5	0.5	80.2	95.5	86.5
Scheme 5	0.1	0.9	**81.5**	**96.4**	**88.4**
Scheme 6	0.1	0.1	80.5	96.0	96.9

Table 7. Ablation studies of the throughput (images/s) for training (Speed) and number of parameters, number of flops in UFDN.

Method	Backbone	Speed (FPS)	Paras (M)	mAP
Baseline	Res50	234	27M	79.6%
Baseline	Swin-T	262	32M	78.2%
CAL	Res50	131	63M	74.3%
TransReID	ViT-B	208	87M	78.0%
UFDN	Res50	161	172M	81.5%
UFDN	Swin-T	150	75M	80.9%

To alleviate the problem, we propose the modulated alignment constraint (MAC) and compare the performance of AC and MAC as shown in Table 5. We speculate that MAC can eliminate the negative impact from the outlier features, and thus experiment baseline+DC+MAC outperforms experiment baseline+DC+AC on two kinds of backbones.

2) Hyperparameters in CDC. During the design of the alignment constraint L_{align}, we validate the influence of different hyper-parameters on the final performance as shown in Table 6. Firstly, we use a margin θ in alignment constraint to control the distance between the positive distance and the negative distance, and the margin ranges between 0 and 1 since that we use Cosine distance. We experiment in Scheme 1-4 in Table 6 and find that when we achieve the best performance when $\theta = 0.1$.

Moreover, we also experiment on the memory bank parameter α which controls the smooth process of the cluster centers in Table 6 (Scheme 2,5,6). We find that a large $\alpha = 0.9$ drives the cluster centers to converge more stable and the best ReID performance.

Ablation Study on Time and Computation. As mentioned before, we decouple the features by adding an extra transformer-based decomposing head and a decouple constraint and then examine the additional time and computation cost in Table 7. Although the THD and CDC modules bring in some extra computation and time cost, UFDN has better trade-off between speed and accuracy when compared with SoTA methods like CAL or TransReID. The model size of UFDN can be further compressed by reducing feature dimension.

5 Visualizations and Discussion

Visualization of UFDN. Fig. 4 shows how the proposed UFDN decouples the Vehicle feature into four parts. The first column of images are the raw images for Vehicle ReID and the right four columns of images represent the gradient-based class activation [30] of different decoupled features. We find the information contained in the decoupled parts are diverse and unstructured, where column (b) pays more attention to the lights of vehicles, column (c) pays more attention

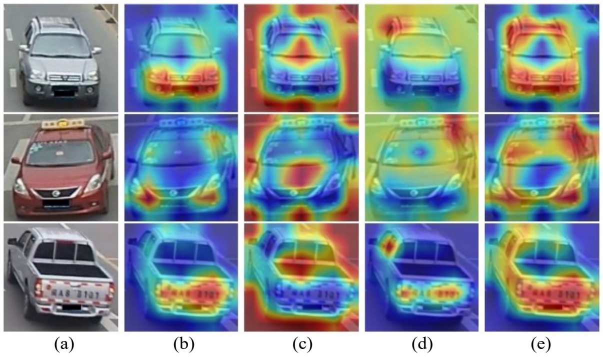

Fig. 4. The Visualization of attention map in reid branch.

to the background and counter information, column (d) focus more attention on the upper information and window information of vehicles, and column (e) focuses more attention on the global information such as colors, types.

Moreover, we also utilize T-SNE [23] to visualize different groups of decoupled features in Fig 5. The raw features (w/o CDC) and the decoupled features (w/ CDC) are visualized in Fig. 5(a) and Fig. 5(b), respectively. The results show that the decoupled features within the same group are clustered together and the decomposed features without CDC are indistinguishable. The above phenomenon validates the effectiveness of our cluster-based decouple constraint.

Comparison with the CNN-based Person ReID Methods. Considering that Vehicle ReID comes under the larger object ReID, we compare with some person ReID methods [16,2] which focus on a similar problem with us. Although these methods [16,2] try to enrich the diversity of features by the diversity constraint, there still exist several differences: 1) we perform the diverse constraint based on the one-dimensional decoupled features that contain more semantic information, but the methods [16,2] are all operated on the two-dimensional feature maps that are local specified and low-level. 2) UFDN employs the self-attention (transformer) mechanism, which uses the extra decomposing token to learn the decoupled features, and thus the learning process of different decoupled parts are independent. However, the CNN-based methods [16,2] are limited by the reception field from the CNN, and different regions or channels tend to have a shared region of interest. 3) UFDN further proposes the alignment constraint for aligning the diverse features which can bring further performance improvement as shown in Table 5. 4) We also compare UFDN with ABDNet [2] (which is open-sourced) on VeRi-776, *e.g.*, 81.5% (UFDN) VS 80.8% (ABDNet) with the same backbone (ResNet50), which shows the priority of our method.

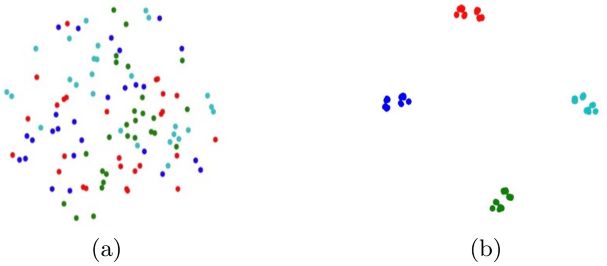

Fig. 5. Tsne visualization of a) the raw features (without decoupling constraint), b) the decoupled features (with decoupling constraint). We decouple the vehicle feature into four parts, where different color represents four groups of decoupled features.

Table 8. Comparison with ABDNet on VeRi-776, where 'W-norm', 'Div' and 'Align' denote weight orthogonality regularizers, the diversity constraint and the alignment constraint, respecitively.

Method	Backbone	Diverse-method	mAP	CMC@1
ABDNet [2]	ResNet50	Div+W-norm	80.8%	96.1%
UFDN	ResNet50	Div	81.2%	96.2%
UFDN	ResNet50	Div+Align	81.5%	96.4%

6 Conclusions

In this paper, we introduce the unstructured feature decoupling network (UFDN) that aims to decouple the vehicle features into unstructured parts and align them without extra annotation. The UFDN consists of a transformer-based feature decomposing head (TDH) and a cluster-based decouple constraint (CDC). We evaluate our UFDN on three popular benchmarks (VeRi-776, VehicleID, and VERI-WILD) and two backbones (ResNet50 and Swin-Tiny) and achieve competitive results when comparing with other works. Firstly, the improvements from the TDH demonstrate that the self-attention mechanism can be used to encode the discriminative information of channel-wise decomposed features into the final decomposing tokens. Secondly, the CDC can force the relative decomposed features from different images to have a similar region of interest. Finally, the improvement achieved by UFDN proves that the decompose and decouple processes in UFDN are effective and can lead to performance progress.

Acknowledgements

This work was supported by the National Science Foundation of China under Grant NSFC 61906194 and the National Key R&D Program of China under Grant 2021YFF0602101. This work was supported by Alibaba Group through Alibaba Research Intern Program.

References

1. Chen, H., Lagadec, B., Bremond, F.: Partition and reunion: A two-branch neural network for vehicle re-identification. In: CVPR Workshops. pp. 184–192 (2019)
2. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8351–8361 (2019)
3. Chen, T., Liu, C., Wu, C., Chien, S.: Orientation-aware vehicle re-identification with semantics-guided part attention network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12347, pp. 330–346. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_20, https://doi.org/10.1007/978-3-030-58536-5_20
4. Chen, Y., Jing, L., Vahdani, E., Zhang, L., He, M., Tian, Y.: Multi-camera vehicle tracking and re-identification on ai city challenge 2019. In: CVPR Workshops. vol. 2 (2019)
5. Guo, H., Zhao, C., Liu, Z., Wang, J., Lu, H.: Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
6. Guo, H., Zhao, C., Liu, Z., Wang, J., Lu, H.: Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: McIlraith, S.A., Weinberger, K.Q. (eds.) AAAI. pp. 6853–6860. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16206
7. He, B., Li, J., Zhao, Y., Tian, Y.: Part-regularized near-duplicate vehicle re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3997–4005 (2019)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90, https://doi.org/10.1109/CVPR.2016.90
10. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
11. Khamis, S., Kuo, C.H., Singh, V.K., Shet, V.D., Davis, L.S.: Joint learning for attribute-consistent person re-identification. In: European Conference on Computer Vision. pp. 134–146. Springer (2014)
12. Khorramshahi, P., Kumar, A., Peri, N., Rambhatla, S.S., Chen, J.C., Chellappa, R.: A dual-path model with adaptive attention for vehicle re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6132–6141 (2019)
13. Khorramshahi, P., Peri, N., Chen, J., Chellappa, R.: The devil is in the details: Self-supervised attention for vehicle re-identification. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV. Lecture Notes in Computer Science, vol. 12359, pp. 369–386. Springer (2020). https://doi.org/10.1007/978-3-030-58568-6_22, https://doi.org/10.1007/978-3-030-58568-6_22

14. Khorramshahi, P., Rambhatla, S.S., Chellappa, R.: Towards accurate visual and natural language-based vehicle retrieval systems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4183–4192 (June 2021)
15. Li, M., Huang, X., Zhang, Z.: Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 194–204 (2021)
16. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 369–378 (2018)
17. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., Yang, Y.: Improving person re-identification by attribute and identity learning. Pattern Recognition **95**, 151–161 (2019)
18. Liu, H., Tian, Y., Yang, Y., Pang, L., Huang, T.: Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2167–2175 (2016)
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
20. Liu, X., Liu, W., Mei, T., Ma, H.: Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia **20**(3), 645–658 (2017)
21. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. CoRR **abs/1711.05101** (2017), `http://arxiv.org/abs/1711.05101`
22. Lou, Y., Bai, Y., Liu, J., Wang, S., Duan, L.: Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3235–3243 (2019)
23. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
24. Meng, D., Li, L., Liu, X., Li, Y., Yang, S., Zha, Z.J., Gao, X., Wang, S., Huang, Q.: Parsing-based view-aware embedding network for vehicle re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7103–7112 (2020)
25. Mo, W., Lv, J.: Cascaded hierarchical context-aware vehicle re-identification. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)
26. Qian, J., Jiang, W., Luo, H., Yu, H.: Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. Measurement Science and Technology (2020)
27. Qian, W., He, Z., Peng, S., Chen, C., Wu, W.: Pseudo graph convolutional network for vehicle reid. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 3162–3171. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475462, `https://doi.org/10.1145/3474085.3475462`
28. Qian, W., Yang, X., Peng, S., Yan, J., Guo, Y.: Learning modulated loss for rotated object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2458–2466 (2021)
29. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1025–1034 (2021)

30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
31. Shen, F., Xie, Y., Zhu, J., Zhu, X., Zeng, H.: Git: Graph interactive transformer for vehicle re-identification. arXiv preprint arXiv:2107.05475 (2021)
32. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
33. Suprem, A., Pu, C.: Looking glamorous: Vehicle re-id in heterogeneous cameras networks with global and local attention. arXiv preprint arXiv:2002.02256 (2020)
34. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
35. Wang, H., Peng, J., Chen, D., Jiang, G., Zhao, T., Fu, X.: Attribute-guided feature learning network for vehicle re-identification. arXiv preprint arXiv:2001.03872 (2020)
36. Wang, H., Peng, J., Jiang, G., Xu, F., Fu, X.: Discriminative feature and dictionary learning with part-aware model for vehicle re-identification. Neurocomputing **438**, 55–62 (2021)
37. Zhang, X., Zhang, R., Cao, J., Gong, D., You, M., Shen, C.: Part-guided attention learning for vehicle re-identification. CoRR **abs/1909.06023** (2019), `http://arxiv.org/abs/1909.06023`
38. Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3186–3195 (2020)
39. Zhao, Y., Shen, C., Wang, H., Chen, S.: Structural analysis of attributes for vehicle re-identification and retrieval. IEEE Transactions on Intelligent Transportation Systems **21**(2), 723–734 (2019)
40. Zheng, A., Lin, X., Li, C., He, R., Tang, J.: Attributes guided feature learning for vehicle re-identification. arXiv preprint arXiv:1905.08997 (2019)
41. Zhou, J., Su, B., Wu, Y.: Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2909–2918 (2020)
42. Zhou, Y., Shao, L.: Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6489–6498. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00679, `http://openaccess.thecvf.com/content_cvpr_2018/html/Zhou_Viewpoint-Aware_Attentive_Multi-View_CVPR_2018_paper.html`
43. Zhu, J., Zeng, H., Huang, J., Liao, S., Lei, Z., Cai, C., Zheng, L.: Vehicle re-identification using quadruple directional deep learning features. IEEE Transactions on Intelligent Transportation Systems **21**(1), 410–420 (2019)
44. Zhu, R., Fang, J., Xu, H., Yu, H., Xue, J.: Dcdlearn: Multi-order deep cross-distance learning for vehicle re-identification. arXiv preprint arXiv:2003.11315 (2020)