MsKAT: Multi-Scale Knowledge-Aware Transformer for Vehicle Re-Identification

Hongchao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo

***Abstract*— Existing vehicle re-identification (Re-ID) methods usually suffer from intra-instance discrepancy and inter-instance similarity. The key to solving this problem lies in filtering out identity-irrelevant interference and collecting identity-relevant vehicle details. In this paper, we aim to design a robust vehicle Re-ID framework that trains a model guided by knowledge vectors yet is able to disentangle the identity-relevant features and identity-irrelevant features. Toward this end, we propose a novel Multi-scale Knowledge-Aware Transformer (MsKAT) to build a knowledge-guided multi-scale feature alignment framework. First, we construct a Knowledge-Aware Transformer (KAT) to interact with semantic knowledge and visual feature. KAT mainly includes State elimination Transformer (SeT) to eliminate state (camera, viewpoint) interference and Attribute aggregation Transformer (AaT) to gather attribute (color, type) information. Second, to learn the knowledge-guided sample differences, we propose to encourage the separation of identity-relevant features and identity-irrelevant features by a Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA}$). Specifically, $\mathcal{L}_{KGA}$ suppresses the difference between knowledge-guided positive pairs and the similarity between knowledge-guided negative pairs. Third, with the multi-scale settings of KAT and $\mathcal{L}_{KGA}$, our model can capture knowledge-guided visual consistency features at different scales. Extensive evidence demonstrates our approach achieves new state-of-the-art on three widely-used vehicle re-identification benchmarks.**

***Index Terms*— Vehicle re-identification, knowledge-aware, transformer, multi-scale.**

I. Introduction

VEHICLE re-identification (Re-ID), which aims to identify vehicle images from the gallery that shares the same vehicle as the given probe, is an active task driven by the applications of smart city and intelligent transportation. Despite years of extensive efforts, it still faces two severe challenges. 1) The intra-instance discrepancy among the same vehicle images under different states, *e.g.*, different camera views, vehicle viewpoints, and capture times. 2) The inter-instance similarity among different vehicles, especially when sharing the same attributes, *e.g.*, the same color, type, and manufacturer.

Recent efforts have provided various solutions while handling the above challenges. Representative approaches fall into three categories: 1) Generation-based methods [1]–[5], which aim to handle viewpoint changes and generate cross-view features to supplement the original features. Methods of this category show one major benefit, which is learning cross-view features to reduce intra-instance differences. However, the unrealistic samples affect the explicit regularization of the feature representations in cross-view generalization. 2) Part-based methods [6]–[11] learn local features to enhance the discriminative clues of global features. Methods of this category show two major benefits: strengthening discriminative regions for distinguishing subtle differences, and aligning parts of cross-view samples for the same identity. However, part extraction model usually requires a large amount of annotated data which is time and labor-consuming. Furthermore, the performance of forthcoming vehicle Re-ID is very sensitive to the inaccurate results of part extraction. 3) Knowledge-based methods [12]–[17] use additional color, type, camera, viewpoint and other prior knowledge to assist the vehicle Re-ID task. Methods of this category show two major benefits: introducing identity-relevant attribute knowledge (*e.g.*, color and type) for global features, and reducing identity-irrelevant state (*e.g.*, camera and viewpoint) changes of hard positive samples for the same identity. In summary, generation-based methods focus on learning identity invariant features in different states, while part-based methods focus on subtle identity-related discriminative clues. However, the performance of existing knowledge-based methods is eclipsed by generation-based methods and part-based methods. The reason is that existing knowledge-based methods still face the following two shortcomings. 1) How to effectively interact with knowledge vectors and feature maps. Existing methods tend to directly cascade semantic information and visual information, ignoring the response relationship between knowledge vector and feature map. 2) How to distinguish different samples under similar knowledge vectors. The common way is to directly pass the cascaded knowledge features into the final loss function, ignoring the knowledge-guided sample differences.

Manuscript received January 6, 2022; accepted March 30, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976002, in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2020-051 and Grant GXXT-2020-013, in part by the Key Project of Research and Development of Anhui Province under Grant 202104d07020008, and in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2020A0033. The Associate Editor for this article was Y. Song. *(Corresponding author: Aihua Zheng.)*

Hongchao Li, Jin Tang, and Bin Luo are with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: lhc950304@foxmail.com; tangjin@ahu.edu.cn; ahu_lb@163.com).

Chenglong Li is with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230601, China, and also with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230601, China (e-mail: lcl1314@foxmail.com).

Aihua Zheng is with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei 230026, China (e-mail: ahzheng214@foxmail.com).

Digital Object Identifier 10.1109/TITS.2022.3166463

To explore the response relationship between knowledge vector and visual feature map, we propose to introduce the transformer [18] architecture into vehicle Re-ID. Transformer has shown its strong ability in modeling the dependence of patches via a self-attention manner [19]. Then, the transformer-based computer vision models [20]–[22] also occupy the top-k ranks on many benchmarks and tasks. However, existing works rarely consider the dependency between visual patches and semantic patches. Therefore, transformer-based visual-semantic interaction is still an interesting problem to be further studied. Different from above transformer-based models, we propose a Knowledge-Aware Transformer (KAT) to interact with semantic knowledge and visual feature. KAT mainly includes State elimination Transformer (SeT) and Attribute aggregation Transformer (AaT) as shown in Fig. 1 (Middle). In vehicle Re-ID, $camera$ and $viewpoint$ variations are two key factors causing the intra-instance discrepancy. We hope vehicle Re-ID system with capability of recognizing the same vehicle captured in different states. Specially, we propose the SeT to integrate the state knowledge into the feature tensor and suppress patches with large similarities, to reduce the interference of camera/viewpoint changes on vehicle features during the feature learning. Meanwhile, $color$ and $type$ similarity are the two key factors causing the inter-instance similarity in vehicle Re-ID. In the same manner, we propose the AaT to collect patches with large similarities of the attribute knowledge and the feature tensor, to learn the color/type nuances of vehicles. However, SeT and AaT result in the loss of some discriminative information, because this process only conveys knowledge clues in a single feature map, ignoring the learning of identity-relevant information and identity-irrelevant information between sample pairs.

Furthermore, to learn the knowledge-guided sample differences, we propose a Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA}$) to learn the identity-relevant features and the identity-irrelevant features respectively. $\mathcal{L}_{KGA}$ mainly constrains knowledge-guided positive pairs and knowledge-guided negative pairs as shown in Fig. 1 (Bottom). Vehicle images from the same category have consistent attribute information and changeable state information. We propose positive knowledge-guided alignment loss ($\mathcal{L}_{KGA_p}$) to encourage the visual features of attribute-aggregation to be aligned and the visual features of state-elimination to be inconsistent. It means that our $\mathcal{L}_{KGA_p}$ reduces the intra-instance differences by discarding identity-irrelevant features in the positive space. Vehicle images from hard negative pairs tend to share consistent attribute information. We propose negative knowledge-guided alignment loss ($\mathcal{L}_{KGA_n}$) to force the visual features of attribute-aggregation to be misaligned, which digs out vehicle details in the same attribute space. $\mathcal{L}_{KGA_p}$ and $\mathcal{L}_{KGA_n}$ can separate identity-related information and identity-irrelevant information in the feature map and are not limited by the size of the scale, it inspires us to do feature alignment learning on different scales.

Naturally, we introduce the multi-scale learning framework into vehicle Re-ID. It has been demonstrated that multi-scale feature learning [23]–[25] improves the capacity of deep networks in image classification, object detection and semantic segmentation. However, these multi-scale feature learning methods tend to interact with a single input image at different scales. We propose to learn the knowledge-guided sample differences at different scales and build a knowledge-guided multi-scale feature alignment framework as shown in Fig. 2 (b). Our goal is to spread the information of different samples at the same scale under the guidance of the knowledge.

The contributions of this paper can be summarized as follows.

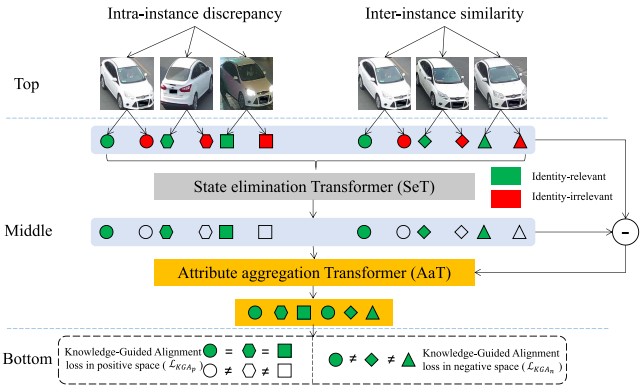

Fig. 1. Illustration of knowledge-aware transformer (KAT) and knowledge-guided alignment loss ($\mathcal{L}_{KGA}$). Vehicle images captured from different states present appearance variations which result in intra-instance discrepancy. Different vehicles that share the same attributes present a similar appearance which results in inter-instance similarity. KAT is designed with a state elimination transformer (SeT) to alleviate state interference and an attribute aggregation transformer (AaT) to aggregate attribute details. To learn the knowledge-guided sample differences, we further use $\mathcal{L}_{KGA_p}$ to learn the consistent features between knowledge-guided positive pairs and use $\mathcal{L}_{KGA_n}$ to learn the discriminative features between knowledge-guided negative pairs.

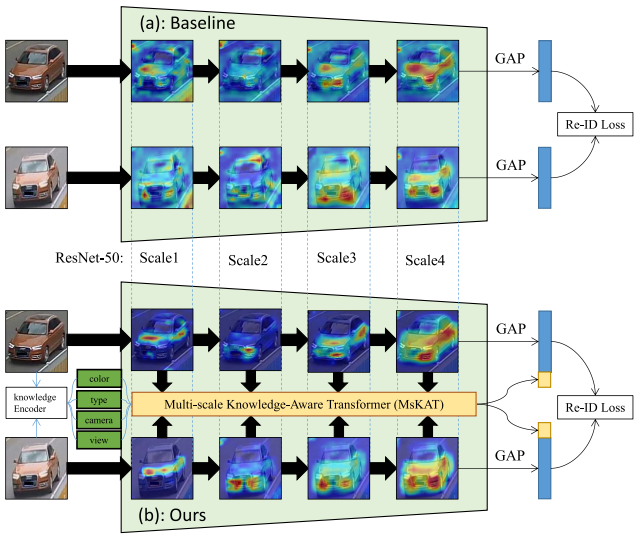

Fig. 2. Visualization of class activation maps at different scales. (a) Baseline: The traditional convolutional network can only constrain the similarity of the feature vectors in the last layer of the network. (b) Ours: We propose a Multi-scale knowledge-aware transformer (MsKAT) to focus on the feature learning of same region at different scales.

- We propose to introduce the transformer structure as an interactive bridge between visual features and semantic knowledge. Particularly, we propose a Knowledge-Aware Transformer (KAT), which eliminates the interference of state information by State elimination Transformer (SeT) and collects attribute information by Attribute aggregation Transformer (AaT).
- We propose a Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA}$) to learn the knowledge-guided sample differences. $\mathcal{L}_{KGA}$ mainly includes $\mathcal{L}_{KGA_p}$ to suppress the difference between knowledge-guided positive pairs and $\mathcal{L}_{KGA_n}$ to reduce the similarity between knowledge-guided negative pairs.
- We propose to build a knowledge-guided multi-scale feature alignment framework, which is guided by KAT and $\mathcal{L}_{KGA}$ at different scales. Existing multi-scale feature learning methods tend to interact with a single input image at different scales. Our method aims to do feature alignment learning on different scales under the guidance of the knowledge.
- Comprehensive experiments on three large-scale vehicle Re-ID benchmark datasets confirm the effectiveness of the proposed model. In addition, sufficient experiments verify the complementarity and effectiveness of each component we proposed.

II. Related Work

We briefly review the related works in the following two folds, *i*.e., vehicle Re-ID and transformer.

A. Vehicle Re-Identification

Due to wide applications in video surveillance and social security, the vehicle Re-ID task has gained more and more attention in recent years. These previous methods can be summarized into three categories:

1) Generation-Based Methods: Methods of this category are mainly to generate cross-view or multi-view features to handle the viewpoint variation issue in vehicle Re-ID, Sochor *et al.* [26] learn a 3D orientation vector embedded into the feature map for vehicle recognition. They show that orientation information can decrease classification error and boost verification average precision. Zhou *et al.* [27] generate the opposite side features to handle the viewpoint problem. Zhou *et al.* [4] propose a viewpoint aware network that integrates features from viewpoint-based feature extractors with a GAN to create cross-view features for vehicle Re-ID. Zhou *et al.* [1] exploit the great advantages of DCNN and Long Short-Term Memory (LSTM) [28] to learn transformations across different viewpoints of vehicles. Lou *et al.* [5] propose an embedding adversarial learning network (EALN) to generate hard negative cross-view and same-view images for more robust training in vehicle Re-ID. Jin *et al.* [2] propose an Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network to exploit the comprehensive information of multi-view of the same vehicle for effective vehicle Re-ID. However, generating cross-view features is unstable and insufficient, and methods of this category always ignore the challenge of inter-instance similarities.

2) Part-Based Methods: Part-based methods utilize discriminative regional features as a complement to the global backbone features. He *et al.* [7] investigate vehicle local regions to learn part-regularized features for vehicle Re-ID. Khorramshahi *et al.* [8] present a dual-path adaptive attention model, to capture key-points related to parts for vehicle Re-ID. Meng *et al.* [29] detected multiple part regions for each vehicle through a U-Net part parser to generate discriminative features. Meng *et al.* [9] propose a part perspective transformation on feature space to transform the deformed region to a unified perspective. Liu *et al.* [11] adopt the graph convolutional networks (GCNs) [30] to model the correlation among parts for vehicle Re-ID. However, the part-based approaches need additional part annotations, which takes extra costs. A part prediction network is also needed, which involves more training procedures and complicates the feature extraction model. In addition, identity-relevant part information is easily disturbed by the challenge of intra-instance differences.

3) Knowledge-Based Methods: The knowledge information, such as spatial-temporal information, vehicle attribute, are aggregated into global vehicle embedding. Liu *et al.* [31] fuse color, texture, and deep features for vehicle Re-ID. They show that deep features outperform the others and feature fusion improves the Re-ID performance. Yan *et al.* [32] model the relationship of vehicle images as a multi-grain list to discriminate appearance-similar vehicles. By introducing multi-grain relationships, they force the deep model to learn the more discriminative feature between different grains over many images. Shen *et al.* [13] investigate spatial-temporal association for effectively regularizing vehicle Re-ID results. The spatial-temporal information along the candidate path is effectively incorporated to estimate the validness confidence of the path. Li *et al.* [15] propose a deep network architecture guided by meaningful attributes, including vehicle viewpoints, types, and colors, for vehicle Re-ID. Zhao *et al.* [16] collect a new vehicle dataset with 21 classes of structural attributes and proposed a region of interest (ROIs-based) vehicle Re-ID method. In this work, we focus on knowledge-based vehicle Re-ID. On the one hand, state knowledge can be used to deal with the view change problem like generation-based methods. On the other hand, attribute knowledge can be used to dig out the local features like part-based methods. We argue that knowledge-based methods can consider both the intra-instance differences and the inter-instance similarities. However, the existing methods ignore the relationship between the knowledge vector and the feature map, and tend to treat the attribute knowledge as a priori information cascading global feature, or subtract the state similarity in the testing stage.

B. Transformer

As convolutional filter weights are usually fixed after training, they cannot be dynamically adapted to different inputs. Many methods have been proposed to alleviate this problem using non-local filters or self-attention operations [33]–[37]. The basic block in a transformer is the self-attention operation, which aggregates information from the entire input sequence [18]. Many studies also show its effectiveness for

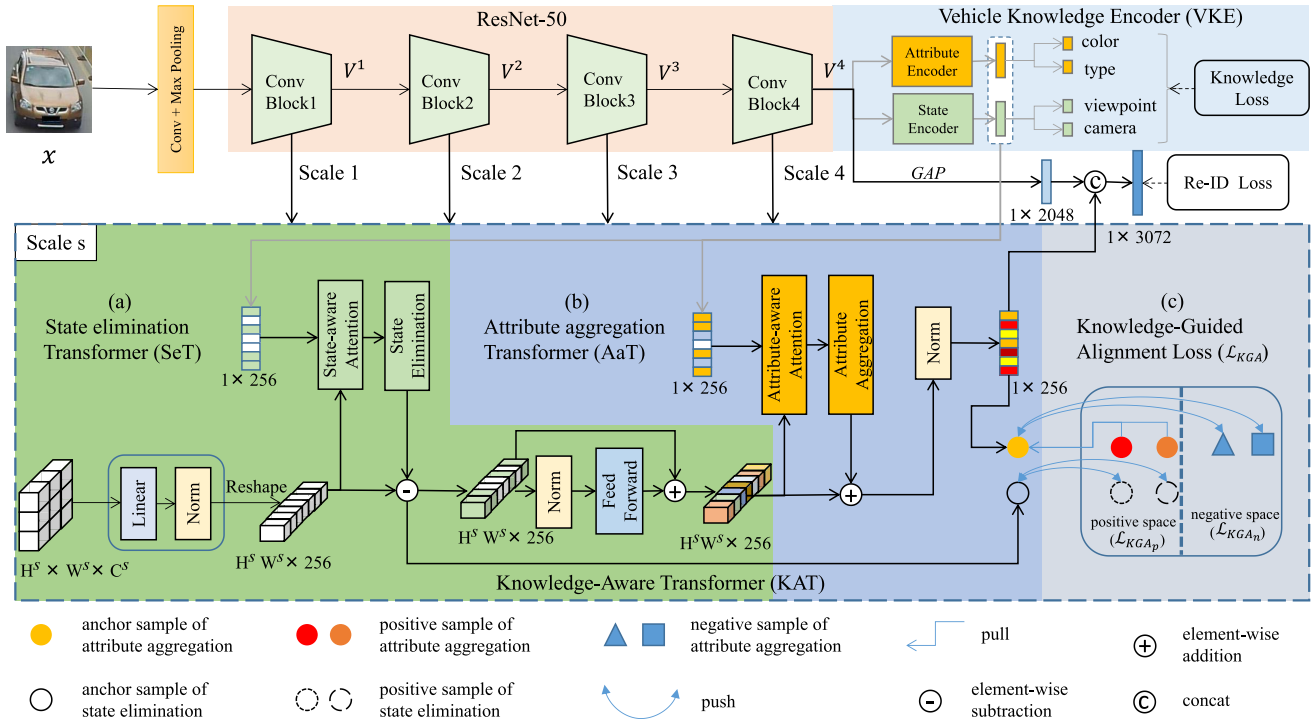

Fig. 3. The architecture of the proposed approach. The whole network consists of a backbone network, vehicle knowledge encoder (VKE), and Knowledge-Aware Transformer(KAT). Specifically, the backbone network uses ResNet-50 pre-trained on ImageNet with the supervision of Re-ID loss to obtain enhanced CNN representation. $V^1 - V^4$ are the feature maps of different scales from the anchor image. VKE employs convolutional blocks with the supervision of knowledge loss to obtain vehicle knowledge vectors. KAT consists of a state elimination transformer (SeT) and an attribute aggregation transformer (AaT). The output of KAT will be divided into identity-relevant features and identity-irrelevant features by knowledge-guided alignment loss ($\mathcal{L}_{KGA}$). The knowledge-aware identity-relevant features at different scales will be used as part of the final features to help vehicle Re-ID. It is worth noting that the knowledge label is not necessary, and the pre-trained knowledge encoder can also be used directly. This is what we did on the vehicleID dataset.

computer-vision tasks. DETR [20] utilizes the transformer decoder to model object detection as an end-to-end dictionary lookup problem with learnable queries, successfully removing the need for handcrafted processes such as NMS. Based on DETR, deformable DETR [38] further adopts a deformable attention layer to focus on a sparse set of contextual elements, obtaining faster convergence and better performance. Recently, vision transformer (ViT) [19] employs a pure transformer model for image classification by treating an image as a sequence of patches. ViT-BoT [21] combines the ViT framework with side information to construct a strong baseline for object re-identification. Later, several studies, such as the T2T-ViT [39], Swin [22], and PVT [40], improved the computation of visual transformers and further boosted their performance. However, existing studies mostly use transformers for feature representation learning, *e.g.* image classification and dense predictions. There lacks a comprehensive study on whether transformers are effective for the interaction between semantic features and visual features. Unlike these approaches, we propose a Knowledge-Aware Transformer architecture to learn identity-relevant features and identity-irrelevant features.

III. Approach

A. Overall Architecture

Our goal is to interact with feature maps and identity-relevant/ -irrelevant knowledge vectors to generate multi-scale knowledge-aware features for robust vehicle Re-ID based on the transformer [18] structure. An overview of our approach is depicted in Fig. 3. It consists of four modules: Backbone network (*e.g.*, ResNet-50), Vehicle Knowledge Encoder (VKE), Knowledge-Aware Transformer (KAT) and the Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA}$). Particularly, KAT is designed as a key module to consider the two major challenges of Re-ID, intra-instance differences and inter-instance similarities. In the KAT module, we first eliminate state discrepancy among intra-instance samples by State elimination Transformer (SeT). Then, an Attribute aggregation Transformer (AaT) is proposed to distill identity-relevant (discriminative) features from those previously reserved by SeT. Moreover, for the KAT module, we design a Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA}$) constraint to optimize two goals: 1) In positive space, we encourage that the features eliminated by the state knowledge to be less discriminative and the features aggregated by attribute knowledge be consistent. 2) In the negative space (especially negative samples with the same attributes), we encourage that the identity-relevant features aggregated by attribute knowledge be inconsistent. Finally, we apply the State elimination Transformer, Attribute aggregation Transformer, and Knowledge-Guided Alignment loss in multiple scales of the deep network, and boost the network to align the identity-relevant information.

B. Backbone Network

State-of-the-art Re-ID methods [9], [11], [29] follow a similar backbone network. They generally train a deep neural network $F(\cdot;\boldsymbol{\theta})$ on the training set, where $\boldsymbol{\theta}$ represents the learnable parameters of the network, and the network is then transferred to extract features from the images in the testing set. Following [9], [11], [29], we adopt ResNet-50 [23] pre-trained on ImageNet [41] as the backbone model in our experiments. We denote a vehicle input as $\boldsymbol{I} = \{(\boldsymbol{x}, y^{id}, y_i^{at}|_{i=1}^{M}, y_j^{st}|_{j=1}^{N})\}$, where $\boldsymbol{x}$ and y^{id} denote the input training vehicle image and its associated vehicle identity label. y_i^{at} and y_j^{st} denote the i-th attribute label and the j-th state label of image $\boldsymbol{x}$ respectively. M and N are the numbers of attribute and state respectively. The corresponding multi-scale feature tensor encoded by the network are denoted as $\boldsymbol{V}^s \in R^{H^s \times W^s \times C^s}, s \in \{1,2,3,4\}$. We use GAP to obtain the vehicle feature vector $\boldsymbol{f} = GAP(\boldsymbol{V}^4) \in R^{C^4}$, where GAP denotes a global average pooling operation. The network parameters $\boldsymbol{\theta}$ is then optimized with respect to a Re-ID loss $\mathcal{L}_{ReID}$ in the form of,

$$\begin{aligned}\mathcal{L}_{ReID} = &-y^{id} log(Softmax(FC(\boldsymbol{f}))) \\ &+ max(0, \|\boldsymbol{f} - \boldsymbol{f}_p)\| + m - \|\boldsymbol{f} - \boldsymbol{f}_n)\|), \end{aligned} \quad (1)$$

where FC denotes a Full Connected layer that predicts the result of classification, $Softmax$ denotes the Softmax function that gets the normalized probability, $\|\cdot\|$ denotes the L_2-norm distance, subscripts p and n indicate the hardest positive and hardest negative feature index in each mini-batch for the sample $\boldsymbol{x}_1$, and $m = 0.3$ denotes the triplet distance margin. $\mathcal{L}_{ReID}$ denotes the widely-used cross-entropy loss [42], and triplet loss [43] with batch hard mining on the Re-ID feature vectors. All samples in the training set will be constrained by the above two loss functions until the training converges. In this paper, we regard ResNet-50 + $\mathcal{L}_{ReID}$ as our baseline. Although this baseline has achieved superior performance in the field of person re-identification [44], it ignores the intra-instance discrepancy and inter-instance similarity between vehicle images. The reason for the intra-instance discrepancy and inter-instance similarity is that vehicles with the same identity are captured in different states (camera, viewpoint), and different vehicles have the same attributes (color, type). We introduce attribute knowledge and state knowledge to disentangle the identity-relevant features and the identity-irrelevant features.

C. Vehicle Knowledge Encoder

Given the vehicle image $\boldsymbol{x}$, we can obtain the feature tensor $\boldsymbol{V}^4$ via the backbone. We design two different convolution blocks to extract two different knowledge vectors respectively:

$$\begin{aligned}\boldsymbol{f}^{at} &= GAP(ReLU(BN(conv_{1\times1}^{at}(\boldsymbol{V}^4)))), \\ \boldsymbol{f}^{st} &= GAP(ReLU(BN(conv_{1\times1}^{st}(\boldsymbol{V}^4)))), \end{aligned} \quad (2)$$

where $conv_{1\times1}^{at}$ and $conv_{1\times1}^{st}$ denote the attribute-related and state-related 1×1 convolutional operation respectivy, BN denotes Bath Normalize operation, $ReLU$ denotes Rectified Linear Unit.

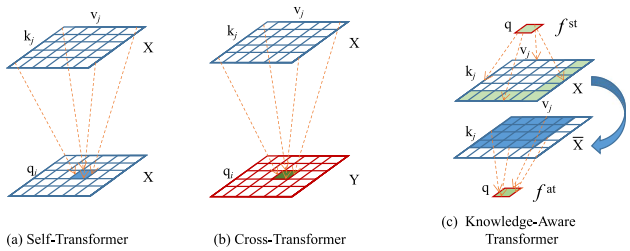

Fig. 4. Self-transformer, cross-transformer and knowledge-aware transformer.

The knowledge vector is constrained by the cross-entropy loss and the ground-truth knowledge label which in the form of,

$$\begin{aligned}\mathcal{L}_{knowledge} = &-\sum_{i=1}^{M} y_i^{at} log(Softmax(FC_i^{at}(\boldsymbol{f}^{at}))) \\ &-\sum_{j=1}^{N} y_j^{st} log(Softmax(FC_j^{st}(\boldsymbol{f}^{st})))), \end{aligned} \quad (3)$$

where FC_i^{at} and FC_j^{st} denote the i-th attribute fully connected layer and the j-th state fully connected layer respectively. M and N are the numbers of attribute and state respectively.

Recent knowledge-based vehicle Re-ID methods [15], [17] have proved that cascading semantic features and visual features are effective for the Re-ID results. However, this interaction strategy of semantic information and visual information is relatively crude, which ignores the response relationship between knowledge vector and feature map. Back to the vehicle Re-ID problem, *camera* and *viewpoint* variations are two key factors causing the intra-instance discrepancy. Meanwhile, *color* and *type* similarity are the two key factors causing the inter-instance similarity. As analyzed above, the Knowledge vector is not just auxiliary information. How to efficiently use state knowledge to eliminate identity-irrelevant information while using attribute knowledge to collect identity-relevant information has became a key issue for knowledge-based vehicle Re-ID methods. It motivates us to find more efficient ways to disentangle identity-relevant features and identity-irrelevant features.

D. Knowledge-Aware Transformer

We propose to introduce the transformer [18] architecture into vehicle Re-ID and convey the rich knowledge cues across feature maps. Transformer has shown its strong ability in modeling the dependence of patches and occupied the top-k ranks on many benchmarks and tasks [20]–[22]. Transformer-based methods have the same core component, i.e., attention mechanism, they can be divided into Self-Transformer and Cross-Transformer. Self-Transformer aims to capture the co-occurring object features on one feature map as shown in Fig. 4 (a). Self-Transformer has been introduced to computer vision such as image classification [19], object detection [20] and object re-identification [21], but it can only capture information on a feature map. Cross-Transformer can exchange

information on different feature maps as shown in Fig. 4 (b). Cross-Transformer is mainly used to complement the texture information of low-resolution images [45] and to fuse information of different modalities [46]. Inspired by the Cross-Transformer, we design a Knowledge-Aware Transformer to explore the potential relevance between knowledge vectors and feature maps.

1) State Elimination Transformer: In the Knowledge-Aware Transformer, we first eliminate state discrepancy among intra-instance samples by State elimination Transformer (SeT). Given an intermediate feature tensor $\boldsymbol{V}^s \in R^{H^s \times W^s \times C^s}$ of height H^s, width W^s, and C^s channels from s-th CNN stage, and a state knowledge vector $\boldsymbol{f}^{st} \in R^c$ of c channels from state encoder, we design a State elimination Transformer, namely SaT, for learning a state-aware attention map of size $H^s \times W^s$. As illustrated in Fig. 3 (a), we scan the spatial positions and assign their patch number as $1, \ldots, H^s W^s$. We take the C-dimensional feature vector at each spatial position as a feature node. We represent the $H^s W^s$ feature nodes as $\boldsymbol{V}_i^s$, where $i = 1, \ldots, H^s W^s$. Generally, state information (camera or viewpoint) is often contained in identity features. Eliminating the state information in different stages can make the learned vehicle features more discriminative.

To facilitate the interaction between the feature tensor and the state knowledge vector at different spatial positions, we introduce transformer structure and first map the feature tensor to the same feature dimension as the state knowledge vector. The query, key, and value of the transformer are expressed as: $\boldsymbol{q} = \boldsymbol{f}^{st}$, $\boldsymbol{k}_i = f_k(\boldsymbol{V}_i^s)$, and $\boldsymbol{v}_i = \boldsymbol{k}_i$, where f_k is embedding function implemented by a linear layer and followed by layer normalization and ReLU activation. Then for each patch k_i in $\boldsymbol{K}$, we calculate the similarity s_i by normalized inner product:

$$s_i = \langle \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|}, \frac{\boldsymbol{k}_i}{\|\boldsymbol{k}_i\|} \rangle. \tag{4}$$

We use the similarity s_i to describe the bi-directional relations between $\boldsymbol{f}^{st}$ and $\boldsymbol{V}_i^s$. Then, we represent the pairwise relations among all the nodes by an affinity matrix $\boldsymbol{S} \in R^{H^s \times W^s}$.

The global feature tensor contains affluent identity-irrelevant information (e.g., camera, viewpoint), we propose to mine state knowledge from them for inferring attention through affinity matrix. We obtain the state-aware attention value $\boldsymbol{f}_i^{set}$ for the $i-$th feature/node through a modeling function as:

$$\boldsymbol{f}_i^{set} = s_i * \boldsymbol{v}_i. \tag{5}$$

The updated feature map can be obtained by aggregating pixel context information of all positions:

$$\boldsymbol{F}^{set} = [\boldsymbol{f}_1^{set}; \boldsymbol{f}_2^{set}; \ldots; \boldsymbol{f}_{H^s \times W^s}^{set}]. \tag{6}$$

SeT reduces state discrepancy and improves anti-interference ability. We propose to restitute the identity-relevant feature to the network by distilling it from the residual feature $\boldsymbol{R}^s$. $\boldsymbol{R}^s$ is defined as

$$\boldsymbol{R}^s = FFN(LN(\boldsymbol{K} - \boldsymbol{F}^{set})). \tag{7}$$

It is also worthy to note that we also utilize the Layer Normalization (LN) and Feed-Forward Networks (FFN) in this procedure.

2) Attribute Aggregation Transformer: Attribute aggregation Transformer (AaT) is proposed to distill identity-relevant (discriminative) features from those previously reserved by SeT. AaT can be categorized as a top-down non-local interaction, which transforms the concept in the higher-level feature maps $\boldsymbol{R}^s$ to the pixels in the lower-level attribute vector $\boldsymbol{f}^{at}$. The output $\boldsymbol{f}^{AaT}$ is a feature vector as $\boldsymbol{f}^{at}$. Same as Eq. (13), we use F_{dot} as the similarity function, which is expressed as:

$$F_{dot}(\boldsymbol{q}, \boldsymbol{k}_j) = \boldsymbol{q}\boldsymbol{k}_j, \tag{8}$$

where $\boldsymbol{q} = \boldsymbol{f}^{at}$ is the attribute vector; $\boldsymbol{k}_j = \boldsymbol{R}_j^s$ is the $j-$th key; $\boldsymbol{v}_j = \boldsymbol{R}_j^s$ is the $j-$th $value$. (k_j) is the $j-$th feature position in $\boldsymbol{R}^s$, then we get the formulation of the proposed AaT as follows:

$$\begin{aligned} &Input : \boldsymbol{q}, \boldsymbol{k}_j, \boldsymbol{v}_j \\ &Similarity : s_j = F_{dot}(\boldsymbol{q}_n, \boldsymbol{k}_j) \\ &Weight : w_j = \frac{exp(s_j)}{\sum_j exp(s_j)} \\ &Output : \boldsymbol{f}^{AaT} = F_{mul}(w_j, \boldsymbol{v}_j), \end{aligned} \tag{9}$$

where F_{mul} is the weight aggregation function (default as $matrix\ multiplication$). Based on Eq. (9), each pair of $\boldsymbol{q}$ and $\boldsymbol{k}_j$ with a closer distance will be given a larger weight. This normalizing function is the standard Softmax.

E. Knowledge-Guided Alignment Loss

To facilitate the disentanglement of identity-relevant features and identity-irrelevant features, we design a Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA}$) constraint by comparing the discrimination capability of features after the SeT and AaT. The core idea is: the identity-relevant features after the attribute aggregation transformer are discriminative, and the identity-irrelevant features after the state elimination transformer are less discriminative. Within a mini-batch, we sample three images, *i.e.*, an anchor sample x, a positive sample x_p, and a negative sample x_n that has a different identity from the anchor sample. For simplicity, we differentiate the two samples by subscript. For example, the feature after attribute aggregation of sample x is denoted by $\boldsymbol{f}^{AaT}$, the feature after state elimination of sample x is denoted by $\boldsymbol{f}^{set} = GAP(\boldsymbol{F}^{set})$. For positive sample pairs, the Knowledge-Guided Alignment loss is thus defined as:

$$\mathcal{L}_{KGA_p} = log(1 + exp(d(\boldsymbol{f}^{AaT}, \boldsymbol{f}_p^{AaT}) - d(\boldsymbol{f}^{SeT}, \boldsymbol{f}_p^{SeT}))), \tag{10}$$

where $d(\boldsymbol{f}, \boldsymbol{f}_p)$ denotes the distance between $\boldsymbol{f}$ and $\boldsymbol{f}_p$ which is defined as $d(\boldsymbol{f}, \boldsymbol{f}_p) = 1 - \frac{\boldsymbol{f}^T \boldsymbol{f}_p}{\|\boldsymbol{f}\| \|\boldsymbol{f}_p\|}$. $log(1 + exp(\cdot))$ is a monotonically increasing function that aims to reduce the optimization difficulty by avoiding negative loss values. For negative sample pairs, the Knowledge-Guided Alignment loss

is defined as:

$$\mathcal{L}_{KGA_n} = log(1 + exp(-d(\boldsymbol{f}^{AaT}, \boldsymbol{f}_n^{AaT}))) * \hat{w}, \tag{11}$$

$$\hat{w} = h(d(\boldsymbol{f}^{at}, \boldsymbol{f}_n^{at}); w) = \begin{cases} 1, & if\ d(\boldsymbol{f}^{at}, \boldsymbol{f}_n^{at}) \leq w, \\ 0, & otherwise, \end{cases} \tag{12}$$

where $\hat{w}$ is a modulation parameter with a value of 0 or 1, which is mainly used to restrict our network to pay more attention to negative samples with high attribute similarity $1 - \frac{\boldsymbol{f}^{at\,T} \boldsymbol{f}_n^{at}}{\|\boldsymbol{f}^{at}\| \|\boldsymbol{f}_n^{at}\|} \leq w$. w is empirically set to 0.2.

On the one hand, the positive Knowledge-Guided Alignment loss $\mathcal{L}_{KGA_p}$ encourages positive samples to be consistent with the information perceived by the knowledge in different feature maps, and forces the positive samples cannot discriminate the features obtained by state elimination. On the other hand, the negative Knowledge-Guided Alignment loss $\mathcal{L}_{KGA_n}$ encourages hard negative samples (which have similar attributes) to be inconsistent with the information perceived by the knowledge in different feature maps. The Knowledge-Guided Alignment loss is $\mathcal{L}_{KGA} = \mathcal{L}_{KGA_p} + \mathcal{L}_{KGA_n}$, rethinking the relationship between positive and negative sample pairs on different scale feature maps from the perspective of knowledge.

1) Overall Loss: We use the commonly used ResNet-50 as backbone network and insert the proposed State elimination Transformer and Attribute aggregation Transformer after each convolution block (in total four convolution blocks)(see Fig. 3). We train the entire network in an end-to-end manner. The overall loss is

$$\mathcal{L} = \mathcal{L}_{ReID} + \mathcal{L}_{knowledge} + \lambda \sum_{s=1}^{n_{scale}} (\mathcal{L}_{KGA}^{s}), \tag{13}$$

where $n_{scale} = 4$ is the number of the stage blocks. λ is a weight that controls the relative importance of the Knowledge-Guided Alignment loss at different hierarchical levels. λ is empirically set to 0.1.

IV. Experience

To validate the superiority of the proposed Multi-scale Knowledge-Aware Transformer (MsKAT) network, it is compared with state-of-the-art vehicle Re-ID approaches on three large-scale databases.

A. Datasets

1) VeRi-776 Dataset: [12] consists of 49357 images of 776 distinct vehicles captured in 20 non-overlapping cameras with various orientations and lighting conditions, where 576 identities with 37778 images and 200 identities with 11579 images are assigned as training and testing respectively. Furthermore, 1678 images from 200 identities have been selected as the queries from the testing set. The original VeRi-776 [12] contains the labels of the vehicle IDs, camera IDs, color IDs and type IDs, while Li *et al.* [15] have annotated the viewpoint information, including $front$, $front\_side$, $side$, $rear\_side$, and $rear$. We use two kinds of state information (camera, viewpoint) and two kinds of attribute information (color, type) in VeRi-776 dataset [12].

2) VERI-Wild Dataset: [47] is a newly released dataset. Different from VeRi-776 [12] captured at day, VERI-Wild [47] are captured at both day and night. The training subset consists of 277797 images of 30671 vehicles. Besides, there are three different scale testing subsets, *i*.e., Test3000 (Small), Test5000 (Middle), and Test10000 (Large). Except for vehicle ID information, VERI-Wild [47] contains various labels of camera, color, type, and manufacturer annotations. Furthermore, we have annotated the time labels according to the acquisition hour of each image. For example, the image captured at 20:19:34 is annotated as 20, and there are 24 time IDs in total. We use two kinds of state information (camera, time) and three kinds of attribute information (color, type, manufacturer) in VERI-Wild dataset [47].

3) VehicleID Dataset: [48] is composed of 221567 images from 26328 unique vehicles. Half of the identities, *i.e.*, 13164, serves for training while the other half for testing evaluation. There are 6 testing splits with various gallery sizes as 800, 1600, 2400, 3200, 6000, and 13164. Following the protocol in [5], [7], [8], we use the first three splits Test800 (Small), Test1600 (Middle) and Test2400 (Large) for testing. During the evaluation, one single image of each identity is randomly selected to form the gallery set while the rest of the images as the query. Which in turn means there is only one ground truth in the gallery for each query image. All the evaluation metrics are based on the average of ten random trials. The vehicle images present in either $front$ or $rear$ viewpoint but without annotation in this dataset. Furthermore, the camera IDs are also not available. Therefore, we use the state and attribute knowledge encoder parameters pre-trained on VERI-Wild [47] to obtain state knowledge and attribute knowledge for VehicleID [48].

B. Evaluation Metrics

Following the general evaluation protocols in the Re-ID field [44], [49], [50], the Rank-1 identification rate (R-1), Rank-5 identification rate (R-5), and mean average precision (mAP) are used as performance metrics. Rank-score is an estimation of finding the correct match in the Rank-K returned results. The mAP is a comprehensive index that considers both the precision and recall of the results. The red, green and blue respectively represent the first, second and third results.

C. Implementation Details

1) Network Architecture: We adopt ResNet-50 [23] pre-trained on ImageNet [41] as the backbone model in our experiments. The classifier weights are randomly initialized. In our implementation, we pad 10 pixels on the image border, and then randomly crop it to 256×256. We normalize RGB channels by subtracting 0.485, 0.456, 0.406 and dividing by 0.229, 0.224, 0.225, respectively. For data augmentation, random erasing is taken to augment the data. Follow [49], we remove the last spatial down-sampling operation in ResNet-50 [23]. The Adam optimizer [56] is used with a batch size of 64 (16 IDs, 4 instances). We run our experiments on two NVIDIA GeForce RTX 2080Ti with 11GB RAM. Follow [44], we use warmup [57] to bootstrap the network, which spent 10 epochs

TABLE I

COMPARISON RESULTS OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VERI-776 DATASET

	Methods	mAP	Rank-1	Rank-5	Reference
(1)	FACT [31]	0.188	0.522	0.729	ICME 2016
	OIFE [51]	0.480	0.659	-	ICCV 2017
	SCPL [13]	0.583	0.835	0.900	ICCV 2017
	NuFACT [14]	0.485	0.769	0.914	TMM 2018
	DF-CVTC [15]	0.611	0.913	0.958	TETCI 2021
	SAN [17]	0.725	0.933	0.971	MST 2020
(2)	GSTE [52]	0.578	0.958	0.965	TMM 2018
	VAMI [4]	0.501	0.770	0.908	CVPR 2018
	EALN [5]	0.574	0.844	0.941	TIP 2019
	FDA-Net [47]	0.555	0.843	0.924	CVPR 2019
	QD-DLF [53]	0.618	0.885	0.945	TITS 2020
	UMTS [2]	0.759	0.958	-	AAAI 2020
(3)	RAM [6]	0.615	0.886	0.940	ICME 2018
	AAVER [8]	0.612	0.890	0.947	ICCV 2019
	PRN [7]	0.743	0.943	0.989	CVPR 2019
	PVEN [29]	0.795	0.956	0.984	CVPR 2020
	SAVER [10]	0.796	0.964	0.986	ECCV 2020
	TBE-Net [54]	0.795	0.960	0.985	TITS 2021
	HPGN [55]	0.802	0.967	-	TITS 2021
	PPT [9]	0.806	0.965	0.983	MM 2020
	Baseline	0.766	0.957	0.980	
	MsKAT	0.820	0.971	0.990	**OURS**

TABLE II

COMPARISON RESULTS OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VEHICLEID DATASET

	Methods	Small		Middle		Large	
		R-1	R-5	R-1	R-5	R-1	R-5
(1)	FACT [31]	0.495	0.680	0.446	0.642	0.399	0.605
	OIFE [51]	-	-	-	-	0.670	0.829
	NuFACT [14]	0.489	0.695	0.436	0.653	0.386	0.607
	DF-CVTC [15]	0.752	0.881	0.722	0.844	0.705	0.821
	ROIVR [16]	0.761	0.912	0.731	0.875	0.712	0.847
	SAN [17]	0.797	0.943	0.784	0.913	0.756	0.883
(2)	GSTE [52]	0.754	0.759	0.743	0.748	0.724	0.740
	VAMI [4]	0.631	0.833	0.529	0.751	0.473	0.703
	FDA-Net [47]	-	-	0.598	0.771	0.555	0.747
	QD-DLF [53]	0.723	0.925	0.707	0.889	0.641	0.834
	EALN [5]	0.751	0.881	0.718	0.839	0.693	0.814
	UMTS [2]	0.809	-	0.788	-	0.761	-
(3)	RAM [6]	0.752	0.915	0.723	0.870	0.677	0.845
	AAVER [8]	0.747	0.938	0.686	0.900	0.635	0.856
	PRN [7]	0.784	0.923	0.750	0.883	0.742	0.864
	PPT [9]	0.796	0.923	0.760	0.894	0.748	0.870
	SAVER [10]	0.799	0.952	0.776	0.911	0.753	0.883
	HPGN [55]	0.839	-	0.800	-	0.773	-
	PVEN [29]	0.847	0.970	0.806	0.945	0.778	0.920
	Baseline	0.802	0.914	0.775	0.887	0.738	0.849
	MsKAT	0.863	0.974	0.818	0.955	0.794	0.939

linearly increasing the learning rate from 3.5×10^{-5} to 3.5×10^{-4}. The learning rate decays to 3.5×10^{-5} and 3.5×10^{-6} at the 40-th epoch and the 70-th epoch respectively. Our model is trained in a total of 120 epochs.

2) Compared Methods: We compare our method with various state-of-the-art methods which mainly fail into three categories.

a) Knowledge-based methods: E.g., Fusion of Attributes and Color feaTures (FACT) [31], Orientation Invariant Feature Embedding (OIFE) [51], Siamese-CNN + Path + LSTM (SCPL) [13], Null space base Fusion of Attribute and Color feaTures (NuFACT) [14], Jointly learns Deep Feature representations, Camera Views, vehicle Types and Colors (DF-CVTC) [15], Two-branch Stripe-based and Attribute-aware Network (SAN) [17], Region of Interests-based Vehicle Re-identification (ROIVR) [16].

b) Generation-based methods: E.g., Viewpoint-aware Attentive Multi-view Inference (VAMI) [4], Group-sensitive Triplet Embedding (GSTE) [52], Embedding Adversarial Learning (EALN) [5], Quadruple Directional Deep Learning Features (QD-DLF) [53], Uncertainty-aware Multi-shot Teacher-Student Network (UMTS) [2], Feature Distance Adversarial Network (FDA-Net) [47], Unlabled-GAN [58].

c) Part-based methods: E.g., Region-aware deep Model (RAM) [6], Adaptive Attention Model for Vehicle Re-identification (AAVER) [8], Part-regularized Near-duplicate (PRN) [7], Part Perspective Transformation (PPT) [9], Hybrid Pyramidal Graph Network (HPGN) [55], Parsing-based View-aware Embedding Network (PVEN) [29], Three-Branch Embedding Network (TBE-Net) [54], t Self-supervised Attention for Vehicle Re-identification (SAVER) [10].

D. Comparison With State-of-the-Art Methods

1) Evaluation Results on VeRi-776: Table I reports the comparison results on VeRi-776 dataset. Part-based methods generally achieve higher performance on VeRi-776 [12] compared with the generation-based methods and the knowledge-based methods. The reason is mainly from the progress of strengthening discriminative regions for distinguishing subtle differences and aligning parts of cross-view samples for the same identity. However, limited by the quality and the number of annotations, the performance of the part-based methods is hard to be further improved and reach a new bottleneck. As shown in Table I, our approach significantly beats the part-based methods as 82.0%, 97.1%, and 99.0% on mAP, the Rank-1, and Rank-5 respectively. Compared with the baseline, our proposed MsKAT significantly improves mAP, Rank-1, and Rank-5 by 5.4%, 1.4%, and 1.0% respectively. This shows the promising achievement by using knowledge to align knowledge-aware information at different scales. Compared with the second-best method PPT [9] with a similar baseline as ours, our approach improves mAP, Rank-1, and Rank-5 by 1.4%, 0.6% and 0.7% respectively. PPT [9] proposes a part perspective transform module to map key points related to part regions to a unified viewpoint on feature space, which only takes into account the alignment of a small number of key points of identity-related information. However, we propose a multi-scale knowledge-aware transformer to eliminate identity-irrelevant information and align identity-relevant information, our MsKAT learns a more robust feature representation on VeRi-776 dataset [12].

2) Evaluation Results on VehicleID: Table II shows the comparison results on VehicleID [48] on three different testing sets. From Table I and Table II, part-based methods does not show the same dominance on the VehicleID [48] dataset as it does on the VeRi776 dataset [12]. For example, the Rank-5 accuracies of SAN [17] achieve 91.3% and 88.3% second only to PVEN [29] as reported in Table II. It means that considering the alignment of the part information is not sufficient for

TABLE III

COMPARISON RESULTS OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VERI-WILD DATASET

	Methods	Small			Middle			Large			Reference
		mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	
(2)	Unlabled-GAN [58]	0.299	0.581	0.796	0.247	0.516	0.744	0.182	0.436	0.655	ICCV 2017
	GSTE [52]	0.314	0.605	0.801	0.262	0.521	0.749	0.195	0.454	0.665	TMM 2018
	FDA-Net [47]	0.351	0.640	0.828	0.298	0.578	0.783	0.228	0.494	0.705	CVPR 2019
	UMTS [2]	0.727	0.845	-	0.661	0.793	-	0.542	0.728	-	AAAI 2020
(3)	AAVER [8]	0.622	0.758	0.927	0.537	0.682	0.889	0.417	0.587	0.876	ICCV 2019
	PPT [9]	0.742	0.919	0.973	0.675	0.891	0.955	0.593	0.848	0.932	MM 2020
	HPGN [55]	0.804	0.914	-	0.752	0.882	-	0.650	0.827	-	TITS 2021
	SAVER [10]	0.809	0.945	0.981	0.753	0.927	0.974	0.677	0.895	0.958	ECCV 2020
	PVEN [29]	0.825	0.967	0.992	0.770	0.954	0.988	0.697	0.934	0.978	CVPR 2020
	Baseline	0.762	0.918	0.966	0.680	0.873	0.945	0.578	0.835	0.917	
	MsKAT	0.840	0.973	0.993	0.787	0.956	0.990	0.722	0.939	0.983	**OURS**

TABLE IV

ABLATION STUDY FOR THE PROPOSED MSKAT

	VKE	KAT		$\mathcal{L}_{KGA}$		VeRi-776		VehicleID						VERI-Wild					
								Small		Middle		Large		Small		Middle		Large	
		SeT	AaT	$\mathcal{L}_{KGA_p}$	$\mathcal{L}_{KGA_n}$	mAP	R-1	R-1	R-5	R-1	R-5	R-1	R-5	mAP	R-1	mAP	R-1	mAP	R-1
a	×	×	×	×	×	0.766	0.957	0.802	0.914	0.775	0.887	0.738	0.849	0.762	0.918	0.680	0.873	0.578	0.835
b	✓	×	×	×	×	0.770	0.957	0.801	0.911	0.768	0.881	0.734	0.842	0.773	0.926	0.703	0.896	0.611	0.877
c	✓	✓	×	×	×	0.776	0.960	0.817	0.928	0.790	0.899	0.755	0.858	0.782	0.935	0.710	0.904	0.628	0.886
d	✓	✓	✓	×	×	0.796	0.967	0.832	0.939	0.782	0.889	0.747	0.862	0.825	0.960	0.772	0.948	0.705	0.932
e	✓	✓	✓	✓	×	0.808	0.970	0.842	0.966	0.810	0.953	0.794	0.936	0.832	0.973	0.780	0.956	0.710	0.939
f	✓	✓	✓	✓	✓	**0.820**	**0.971**	**0.863**	**0.974**	**0.818**	**0.955**	**0.794**	**0.938**	**0.840**	**0.973**	**0.787**	**0.957**	**0.722**	**0.941**

the VehicleID [48], which suffers from drastic viewpoint changes. As shown in Table II, our approach significantly beats the second-best method PVEN [29] by 86.3%, 81.8% and 79.4% on the three different testing sets respectively. Through the effective interaction of knowledge vectors and the consistent learning of knowledge-guided features, our proposed MsKAT approach improves the Rank-1 of three different testing sets by 1.6%, 1.2%, and 1.6% respectively. Note that our method, MsKAT, uses the knowledge encoder previewed on VERI-Wild [47] without any knowledge annotation on VehicleID [48], it further verifies the generality of our method on more general scenarios.

3) Evaluation Results on VERI-Wild: Table III shows the comparison results on VERI-Wild [47] on three different testing sets. As shown in Table III, our MsKAT achieves competitive results on all of the testing subsets on the VERI-Wild dataset [47]. Specifically, the mAP of our method achieves 84.0%, 78.7% and 72.2% on Test3000 (Small), Test5000 (Middle) and Test10000 (Large) respectively, which improves 1.5%, 1.7% and 2.5% than the second-best method PVEN [9]. PVEN [9] proposes a parsing-based view-aware embedding network to achieve the view-aware feature alignment and enhancement for vehicle Re-ID. The PVEN [9] consists of three modules: vehicle part parser, view-aware feature alignment, and common-visible feature enhancement. However, PVEN [9] needs to introduce U-Net as a vehicle part parser and ignores that vehicle parts may be invisible under different states. In comparison, our method only needs one layer of simple convolution to encode the knowledge vector, it implies promising performance in potential large-scale applications.

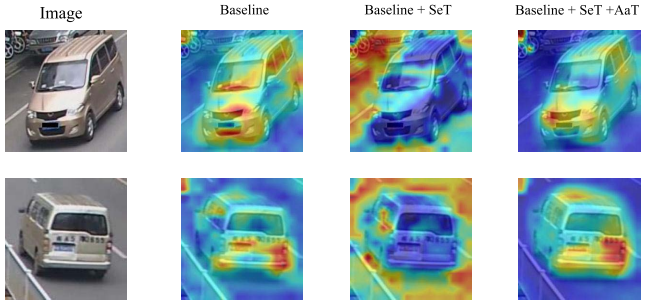

Fig. 5. Activation maps of different features within MsKAT (Conv Block3). They show that SeT focuses on eliminating features that are not related to identity, while AaT focuses on collecting identity-relevant features.

E. Ablation Study

1) Quantitative Analysis of MsKAT: Our baseline is ResNet-50 with $\mathcal{L}_{ReID}$. The Vehicle Knowledge Encoder (VKE) is to facilitates the extraction of knowledge vectors, and the vehicle knowledge encoder with frozen parameters is used in VehicleID [48]. As reported in Table IV (a, b), adding a knowledge encoder to the back of ResNet-50 improves the performance. Especially on the VERI-Wild [47], which implies that the manufacturer knowledge promotes re-identification results. To verify the contribution of State elimination Transformer (SeT) and Attribute aggregation Transformer (AaT) on the three datasets. We progressively introduce the SeT and AaT into the baseline. Both mAP, and Rank-1 scores significantly increase on all the three datasets as shown in Table IV (c, d). However, the current performance is still limited, because the SeT may lose identity-relevant information, and the

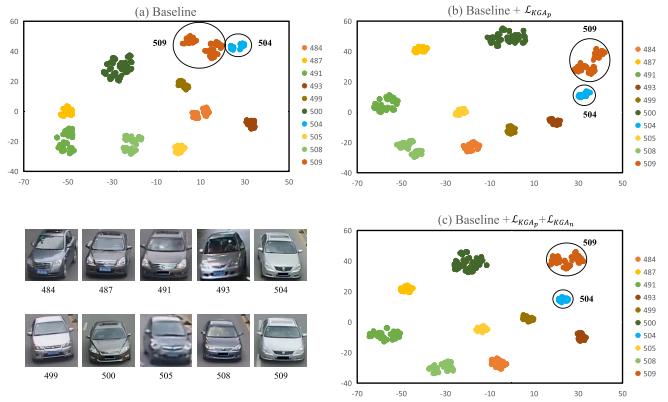

Fig. 6. T-SNE visualization of feature distributions, from *Baseline*, $Baseline + \mathcal{L}_{KGA_p}$ and $Baseline + \mathcal{L}_{KGA_p} + \mathcal{L}_{KGA_n}$. Ten identities with gray color and sedan type are selected from the VeRi-776 and their IDs are listed on the right side of graphs. Circles 504 and 509 contain the features from IDs 504 and 509, respectively.

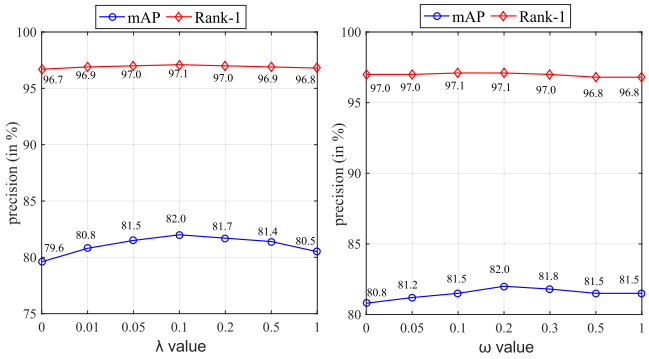

Fig. 7. Parameter analysis at mAP and Rank-1 on VeRi-776.

AaT may collect identity-irrelevant information. We further introduce Knowledge-Guided Alignment loss to disentangle identity-relevant features and identity-irrelevant features and use knowledge to guide visual consistency learning. By adding positive Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA_p}$) and negative Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA_n}$) into the Knowledge-Aware Transformer, both Rank-1 and mAP significantly increase as shown in Table IV (e, f).

2) Qualitative Analysis of MsKAT: To better visualize the contribution of the State elimination Transformer (SeT) and Attribute aggregation Transformer (AaT), We show the visual activation maps as shown in Fig. 5. We can see that SeT mainly eliminates the information of identity-irrelevant areas, and AaT mainly interacts with the information of identity-relevant areas. It means that our proposed SeT and AaT can separate the identity-irrelevant features and the identity-relevant features. To better visualize the contribution of the positive Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA_p}$) and negative Knowledge-Guided Alignment loss ($\mathcal{L}_{KGA_n}$), we demonstrate the T-SNE [59] visualization of feature distribution on VeRi-776 [12] as shown in Fig. 6. The $\mathcal{L}_{KGA_p}$ decreases the intra-class distance of positive samples, while the $\mathcal{L}_{KGA_n}$ increases the inter-class distance of negative samples with the same attribute. It verifies the effectiveness of the Knowledge-Guided Alignment loss, which can reduce the intra-instance discrepancy and increase the inter-instance similarity between vehicle images.

TABLE V

STAGE STUDY OF ADDING KNOWLEDGE-AWARE TRANSFORMER ON VERI-776 AND VERI-WILD

Method	VeRi-776		VERI-Wild-Small	
	mAP	Rank-1	mAP	Rank-1
Baseline	0.766	0.957	0.762	0.918
+ scale 1	0.785	0.963	0.794	0.935
+ scale (1 + 2)	0.802	0.964	0.818	0.949
+ scale (1 + 2 + 3)	0.815	0.967	0.832	0.968
+ scale (1 + 2 + 3 + 4)	**0.820**	**0.971**	**0.840**	**0.973**

TABLE VI

LOSS FUNCTION STUDY OF CHANGING RE-ID LOSS ON VERI-776

Method	mAP	Rank-1	Rank-5
MsKAT (cross-entropy[42])	0.812	0.968	0.985
MsKAT (cross-entropy[42]+circle[60])	0.803	0.968	0.984
MsKAT (cross-entropy[42]+instance[61])	0.817	0.969	0.987
MsKAT (cross-entropy[42]+triplet[43])	**0.820**	**0.971**	**0.990**

F. Other Analysis

1) Design Choices of Knowledge-Aware Transformer: We progressively introduce the Knowledge-Aware Transformer (KAT) to the multiple scales. Both mAP, and Rank-1 scores significantly increase on VeRi-776 [12] and VERI-Wild [47] as shown in Table V. When KAT is added to all four scales, we achieve the best performance. It verifies the effectiveness of our knowledge-guided multi-scale feature alignment framework, which aligns identity-relevant information on multiple scales to boost vehicle Re-ID.

2) Parameter Analysis: There are two important parameters in our model. λ balances the contribution of global feature and cross knowledge-aware feature while ω control the threshold between negative samples constrained by negative Knowledge-Guided Alignment loss. We empirically set $\lambda = 0.1$, $\omega = 0.2$. The parameter analysis results with different λ and ω on VeRi-776 [12] are shown in Fig. 7, which demonstrates that our model is not sensitive to the parameters.

3) Analysis of Different Re-ID Loss Functions: As shown in Eq. (1), we mainly use the widely used cross-entropy loss [42] and triplet loss [43] for network training in this paper. Recently, other loss functions such as circle loss [60] and instance loss [61] have also been applied in Re-ID tasks. To analyze the influence of different Re-ID loss functions on our MsKAT model, we analyze the performance of different loss functions in Table VI. The combination of cross-entropy loss [42] and triplet loss [43] makes our MsKAT model achieve the best performance, which shows that it is effective to disentangle identity-relevant and identity-irrelevant features in hard positive/negative sample pairs.

V. CONCLUSION

In this paper, we propose a Multi-scale Knowledge-Aware Transformer (MsKAT) to build a knowledge-guided

multi-scale feature alignment framework for vehicle Re-ID. First, we introduce Knowledge-Aware Transformer to interact with semantic knowledge and visual features, which eliminates the interference of state (camera, viewpoint) information by State elimination Transformer and collects attribute (color, type) information by Attribute aggregation Transformer. Second, we design a Knowledge-Guided Alignment loss to efficiently disentangle the identity-relevant and identity-irrelevant features, which reduces the difference between positive pairs and the similarity between negative pairs. With the multi-scale settings of Knowledge-Aware Transformer and Knowledge-Guided Alignment loss, our model is able to grasp the stronger aligned details of intra-instance vehicle images. In the future, we will consider building a knowledge graph for vehicles and design the knowledge-guided external transformer for vehicle Re-ID.

REFERENCES

[1] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3278, Mar. 2018, doi: 10.1109/TIP.2018.2819820.

[2] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 11165–11172.

[3] A. Porrello, L. Bergamini, and S. Calderara, "Robust re-identification by multiple views knowledge distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 93–110.

[4] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[5] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3794–3807, Aug. 2019, doi: 10.1109/TIP.2019.2902112.

[6] X. Liu, S. Zhang, Q. Huang, and W. Gao, "RAM: A region-aware deep model for vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[7] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3997–4005.

[8] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6132–6141.

[9] D. Meng, L. Li, S. Wang, X. Gao, Z.-J. Zha, and Q. Huang, "Fine-grained feature alignment with part perspective transformation for vehicle Reid," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 619–627.

[10] P. Khorramshahi, N. Peri, J.-C. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 369–386.

[11] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 907–915.

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

[13] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-ID with Visual-spatio-Temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1918–1927.

[14] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018, doi: 10.1109/TMM.2017.2751966.

[15] H. Li *et al.*, "Attributes guided feature learning for vehicle re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, early access, Dec. 1, 2021, doi: 10.1109/TETCI.2021.3127906.

[16] Y. Zhao, C. Shen, H. Wang, and S. Chen, "Structural analysis of attributes for vehicle re-identification and retrieval," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 2, pp. 723–734, Feb. 2020, doi: 10.1109/TITS.2019.2896273.

[17] J. Qian, W. Jiang, H. Luo, and H. Yu, "Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification," *Meas. Sci. Technol.*, vol. 31, no. 9, Sep. 2020, Art. no. 095401.

[18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–22.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[21] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15013–15022.

[22] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[24] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[25] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[26] J. Sochor, A. Herout, and J. Havel, "Boxcars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3006–3015.

[27] Y. Zhou and L. Shao, "Cross-view GAN based vehicle generation for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] D. Meng *et al.*, "Parsing-based view-aware embedding network for vehicle re-identification," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7103–7112.

[30] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[31] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

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

[33] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2018, pp. 7794–7803.

[34] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[35] L. Wu, R. Hong, Y. Wang, and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2081–2092, Apr. 2020, doi: 10.1109/TCSVT.2019.2909549.

[36] L. Wu, Y. Wang, J. Gao, M. Wang, Z. J. Zha, and D. Tao, "Deep coattention-based comparator for relative representation learning in person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 722–735, Feb. 2021, doi: 10.1109/TNNLS.2020.2979190.

[37] D. Liu, L. Wu, F. Zheng, L. Liu, and M. Wang, "Verbal-person nets: Pose-guided multi-granularity Language-to-Person generation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 9, 2022, doi: 10.1109/TNNLS.2022.3151631.

[38] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–16.

[39] L. Yuan *et al.*, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 558–567.

[40] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[42] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1735–1742.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[44] H. Luo *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020, doi: 10.1109/TMM.2019.2958756.

[45] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5791–5800.

[46] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 214–229.

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

[48] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2167–2175.

[49] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–518.

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[51] Z. Wang *et al.*, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 379–387.

[52] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018. 10.1109/TMM.2018.2796240.

[53] J. Zhu *et al.*, "Vehicle re-identification using quadruple directional deep learning features," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 410–420, Jan. 2020, doi: 10.1109/TITS.2019.2901312.

[54] W. Sun, G. Dai, X. Zhang, X. He, and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, early access, Dec. 3, 2021, doi: 10.1109/TITS.2021.3130403.

[55] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, early access, Jun. 16, 2021, doi: 10.1109/TITS.2021.3086142.

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

[57] X. Fan, W. Jiang, H. Luo, and M. Fei, "SphereReID: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 51–58, Apr. 2019.

[58] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[59] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[60] Y. Sun *et al.*, "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6398–6407.

[61] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, May 2020, doi: 10.1145/3383184.

Hongchao Li received the B.Eng. degree in software engineering from Anhui University, Hefei, China, in 2017, where he is currently pursuing the Ph.D. degree in computer science and technology. His current research interests include person/vehicle re-identification and multi-modal learning.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he worked as a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently an Associate Professor with the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Aihua Zheng received the B.Eng. degree in computer science and technology from the Anhui University of China in 2006, and the Ph.D degree in computer science from the University of Greenwich, U.K., in 2012. She is currently an Associate Professor with the School of Artificial Intelligence, Anhui University. Her main research interests include computer vision and artificial intelligent, especially on person/vehicle re-identification, audio-visual learning, and multi-modal and cross-modal learning.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively. He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.

Bin Luo received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, U.K., in 2002. He is currently a Professor at Anhui University. He is also the chairs of the IEEE Hefei Subsection. He has published more than 200 papers in journals and refereed conferences. He served as a Peer-Reviewer for international academic journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, and *Pattern Recognition Letters*. His current research interests include random graph-based pattern recognition, image and graph matching, and spectral analysis.