This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3283054

Camera Topology Graph Guided Vehicle Re-identification

Hongchao Li, Aihua Zheng, Liping Sun, Yonglong Luo*

***Abstract*—Vehicle re-identification (Re-ID) aims to retrieve vehicles across non-overlapping cameras. Most studies consider representation learning from single appearance information of the vehicle images. Some works adopt the spatio-temporal information to remove unreasonable vehicles to refine the results in the testing phase. However, they ignore the potential topological relations among cameras under the Closed Circuit Television (CCTV) camera systems in the training phase, which usually leads to suboptimal results due to the high intra-identity variations. To handle this problem, we propose a novel vehicle re-identification framework, which explicitly models the camera topological relations of all input images to aggregate neighbor images and thus acquires camera-independent representations. Specifically, we first construct a Camera Topology Graph (CTG) to elucidate the topological relations among cameras. It takes different cameras as nodes and constructs edges from four levels of the camera system, position, orientation, and individual. Then, we introduce a Camera Topology-based Graph Convolutional Network (CT-GCN), which suppresses irrelevant neighbor images and learns different camera representation functions. Finally, we propose a topological cross-entropy loss to obtain the more discriminative vehicle representations. The whole network is trained in an end-to-end manner. Extensive experiments on three benchmark datasets demonstrate the effectiveness of the proposed method against state-of-the-art vehicle Re-ID methods.**

***Index Terms*—Vehicle Re-identification, Closed Circuit Television, Camera Topology Graph, Graph Convolutional Network**

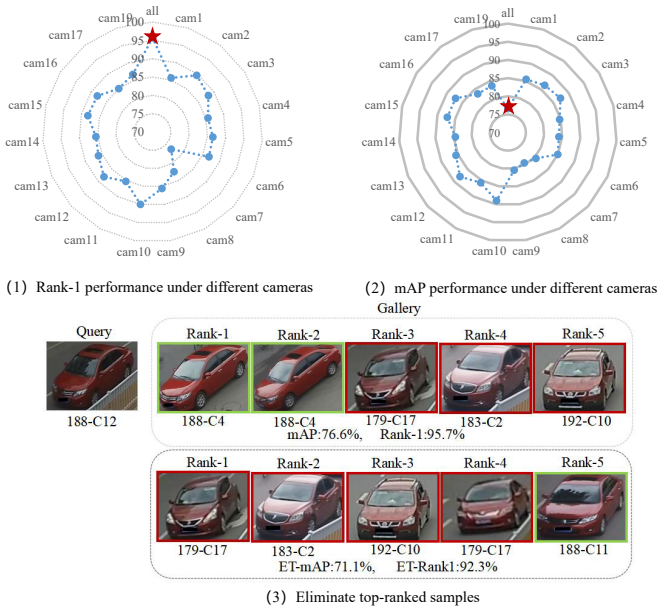


Fig. 1. The phenomena of the strong Re-ID baseline [10] on the VeRi-776 dataset: (1) Rank-1 performance under each camera is lower than performance under all cameras; (2) mAP performance under each camera is higher than performance under all cameras; (3) Eliminating the top-ranked samples (samples from the same camera and the same identity as top-1) significantly degrades the Re-ID performance.

I. Introduction

VEHICLE re-identification (Re-ID) aims to identify vehicle images from the gallery images captured from non-overlapping surveillance cameras that share the same identity as the given probe vehicle. It is an active and challenging task and has drawn more attention due to its wide applications in social security, smart city, and intelligent transportation. Despite the remarkable success, it still faces severe challenges, such as inner-camera occlusions, cross-camera illumination, and viewpoint changes, which restricts its applications in realistic complicated scenarios.

Recently, various efforts have emerged for vehicle Re-ID, including viewpoint-based learning [1], [2], [3], part-based learning [4], [5], [6] and path-based learning [7], [8], [9] and so on. To alleviate viewpoint changes of the vehicle, Zhou *et al.* [1] propose a viewpoint-aware attentive multi-view inference (VAMI) model to infer multi-view features from single-view image inputs. Lou *et al.* [2] propose an embedding adversarial learning network (EALN) to support hard negative and cross-view generation for more robust training in vehicle Re-ID. Chu *et al.* [3] learn two metrics for similar and different viewpoints respectively in two feature spaces and propose a viewpoint-aware network (VANet) for vehicle Re-ID. To further learn local details of vehicles, Khorramshahi *et al.* [4] propose a dual-path adaptive attention model to capture key points related to parts for vehicle Re-ID (AAVER). Meng *et al.* [5] investigate multiple part regions for each vehicle through a U-Net part parser to generate discriminative features. Shen *et al.* [6] adopt the traditional graph convolutional network (GCN) [11] to model the correlation among regions for vehicle

This research is supported in part by the National Natural Science Foundation of China (61976002, 61972439, 62272006), the Natural Science Foundation of Anhui Province (2108085MF214), the Key Program in the Youth Elite Support Plan in Universities of Anhui Province (gxyqZD2020004), and the University Synergy Innovation Program of Anhui Province (GXXT-2021-007).

Hongchao Li, Liping Sun, Yonglong Luo are with Anhui Provincial Key Laboratory of Network and Information Security, School of Computer and Information, Anhui Normal University, Wuhu, 241003, China.

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, 230601, China.

Corresponding author: Yonglong Luo. Email: ylluo@ustc.edu.cn.

Re-ID. Note that the path information[1] is generally available due to the rapid popularization of Closed Circuit Television (CCTV) camera systems, Shen *et al.* [8] investigate spatio-temporal association to estimate the validness confidence of the path. Lv *et al.* [12] construct spatio-temporal constraints and refine the matching results for vehicle Re-ID. Prasad *et al.* [13] use spatio-temporal information as physical constraints to reduce the complexity of the matching algorithm. However, these works mainly focus on mining the information inside a single image and thus lack the interaction between different images.

By analyzing the strong Re-ID baseline [10], we have found the following three phenomena: **Phenomena (1)**: *the Rank-1 performance under the entire camera system is much higher than the performance under each camera*, as shown in Fig. 1 (a). This shows that the Rank-1 performance of the previous methods is inflated, since it tends to retrieve easy positive samples[2] under the entire camera system, and can not accurately hit the positive samples across each camera. **Phenomena (2)**: *the mAP performance under the entire camera system is much lower than the performance under each camera*, as shown in Fig. 1 (b). This shows that the positive samples under each camera are more clustered than those under the entire camera system. **Phenomena (3)**: Eliminating the top-ranked samples significantly degrades the Re-ID performance as shown in Fig. 1 (c). This shows that the Re-ID performance obtained by the conventional methods is suboptimal and susceptible to camera interference.

In addition, we observe that the information of each identity under each camera is limited [14], [15], [16]. Our intuitive solution is to aggregate the information of the vehicle under the entire camera system, thus supplementing vehicle information from different cameras. Therefore, we propose a novel Camera Topology-based Graph Convolutional Network (CT-GCN) for vehicle re-identification to fully explore the easy- and hard-positive samples under the whole camera system and build the bridge between representation models and camera systems.

Concretely, we first employ ResNet-50 [17] to obtain the initial vehicle feature representations. Next, as shown in Fig. 2, we construct the Camera Topology Graph (CTG) to model the relationship of different vehicle images, which is *simple* and *general*. The CTG takes different cameras as nodes and builds edges based on the multiple relationships (*e.g.*, position, orientation) between cameras. The construction details of the CTG are described in Section. III. Once we have the camera topology graph, we argue that we do not need to constrain all positive samples to present the same feature representation, while only interacting with positive samples under neighbor cameras to cover the entire CCTV camera system. For example, given a query image from camera 3 to retrieve the vehicle-of-interest in the gallery set, we can supplement the vehicle information with camera 8 and camera 9 on the driving route of the vehicle, and then retrieve the vehicle-of-interest in the gallery set. As shown in Fig. 2 (c),

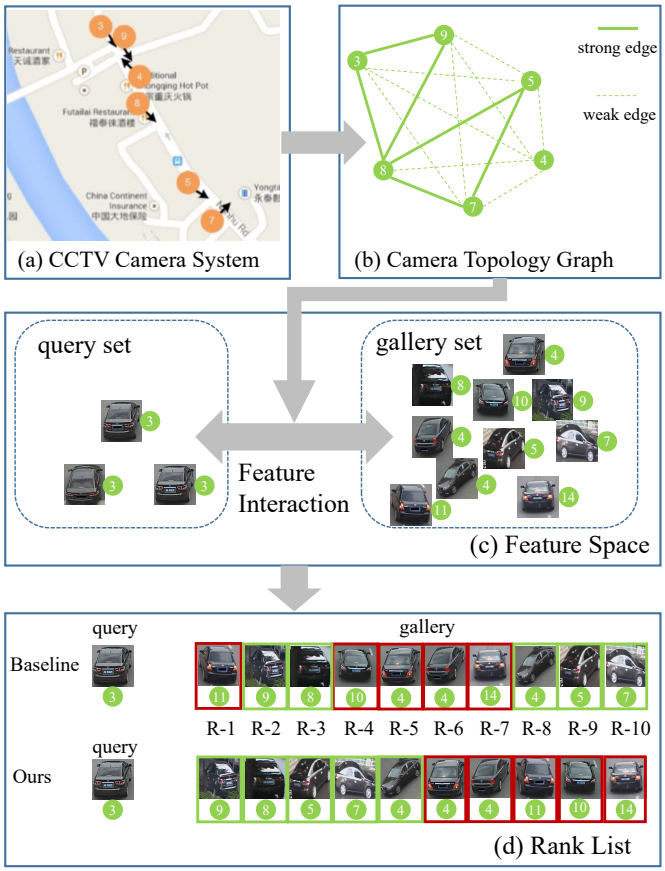


Fig. 2. Illustration of generating a camera topology graph from the real-world traffic scene, a medium for feature interaction of vehicle images. These circled numbers represent camera labels on VeRi-776 dataset [7]. The black arrow indicates the orientation of the camera. Cameras at two consecutive junctions are close in position. There is a strong topological relationship between two cameras only when they are oriented in the same direction and spatially adjacent.

the information from the easy-positive samples[3] is transferred to the hard-positive samples[4] along the camera topology graph, which improves the ranking results. The ranking results as shown in Fig. 2 (d). This means that the proposed camera topology graph meaningfully links the vehicle representation model and the CCTV camera system, which is rarely covered by previous works.

Whereafter, we propose a novel Camera Topology-based Graph Convolutional Network (CT-GCN) to learn the more discriminative cross-camera vehicle feature representations. In a traditional Graph Convolutional Network (GCN) [11], the adjacency matrix of nodes is fixed and each node shares the same weight matrix. The proposed CT-GCN aggregates only the manageable neighbor nodes and learns different weight matrices for different cameras. In general, CT-GCN retains the ability of the traditional GCN to interact with graph nodes, while introducing the learning of different camera topological relations.

[1]Path information refers to the spatio-temporal state of vehicle movement.
[2]The positive samples belong to the same class as the query sample.

[3]For the sample of camera 8, the samples from camera 5/7 can be seen as the easy-positive samples.
[4]For the sample of camera 3, the samples from camera 5/7 can be seen as the hard-positive samples.

Finally, we use the cross-entropy and triple losses to train the whole network in an end-to-end way by following [4], [5], [6]. However, traditional cross-entropy loss can not work with topological relations between positive samples. To optimize the topological relations between positive samples, we propose a novel topological cross-entropy loss to force the network paying more attention to positive samples from neighbor cameras. Such topological cross-entropy loss is also the key to aggregating vehicles under the neighbor cameras, which makes the representation aggregation process more effective and efficient.

Overall, The contributions of this paper can be summarized as follows.

- We design a general way of constructing the camera topology graph to preserve camera topological relations under the real-world traffic scene. This strategy of building a camera topology graph can be expanded to any traffic scene to connect vehicle representation models and CCTV camera systems.
- We propose a novel Camera Topology-based Graph Convolutional Network (CT-GCN) to learn cross-camera vehicle representations. Compared with traditional GCN, CT-GCN has the advantage of adaptively assigning weight matrixes for different cameras and aggregating relevant neighbor nodes.
- We propose a novel topological cross-entropy loss to optimize the feature learning of positive vehicle images. In contrast to conventional cross-entropy loss, we also consider the topological relationships between samples. Our key idea is to preferentially train the easy-positive samples instead of training all positive samples at once.
- Comprehensive experiments on three large-scale vehicle datasets confirm the effectiveness of the proposed framework. Furthermore, sufficient experiments verify the complementarity and effectiveness of each component we proposed.

II. Related Work

We briefly review the related work in the following two folds, i.e., Vehicle Re-identification and Graph Convolutional Network (GCN).

A. Vehicle Re-identification

Because of its wide applications in video surveillance and social security, the task of vehicle Re-ID has earned more and more attention in recent years. Liu *et al.* [18] provide a deep relative distance learning method for measuring the instance difference between different vehicles. Yan *et al.* [19] propose a multi-grain based list ranking (MGLR) approach to build the relationships between vehicle images. Bai *et al.* [20] propose a group-sensitive triplet embedding method to accelerate feature learning and promote the discrimination power for vehicle Re-ID. Different from the above methods exploring the global features, He *et al.* [21] detect windows and lights through a YOLO detector to learn part-regularized features for vehicle Re-ID. Meng *et al.* [5] propose a parsing-based view-aware embedding network (PVEN) to achieve the part alignment and enhancement for vehicle Re-ID. Liu *et al.* [22] propose a group-group loss-based global-regional feature learning framework to optimize the distance within and across vehicle image groups. Li *et al.* [23] propose a multi-scale knowledge-aware transformer (MsKAT) to build a knowledge-guided multi-scale feature alignment framework for vehicle Re-ID. Khorramshahi *et al.* [24] present self-supervised attention for vehicle re-identification (SAVER) to learn vehicle-specific discriminative features. Zhao *et al.* [25] propose a Heterogeneous Relational Complement Network (HRCN) by incorporating region-specific features and cross-level features as complements for the global feature. Li *et al.* [26] propose an attribute and state guided structural embedding network (ASSEN) to achieve discriminative feature learning by attribute-based enhancement and state-based weakening.

To handle the camera variation issue in vehicle Re-ID, Sochor *et al.* [27] learn a 3D orientation vector to decrease classification error and boost verification average precision for vehicle recognition. Zhou *et al.* [28] propose a cross-view generative adversarial network to generate cross-view vehicle images from an input view. Zhou *et al.* [1] adopt a viewpoint-aware attention model and the adversarial training architecture to implement effective multi-view feature inference from single-view input. Lou *et al.* [2] design an embedding adversarial learning network (EALN) into the vehicle Re-ID framework for hard negative and cross-view generation. Chu *et al.* [3] propose two metrics for vehicle re-identification through a viewpoint-aware network. By contrast, path-based methods typically employ spatio-temporal cues to refine the search space. Liu *et al.* [7] propose a spatio-temporal relation model to re-rank vehicles and refine the final results for vehicle Re-ID. Wang *et al.* [9] propose an orientation-invariant regularization to describe the macroscopic embedding of vehicles with constraints on spatio-temporal relationships. Shen *et al.* [8] investigate spatio-temporal association for effectively refining vehicle Re-ID results. Lv *et al.* [12] construct spatio-temporal constraints and refine the retrieving results for vehicle Re-ID. Prasad *et al.* [13] use spatio-temporal information as physical constraints to reduce the complexity of the retrieving algorithm. Although these methods have made great progress in the Re-ID problem, they ignore potential topological relations of vehicle images under the CCTV camera systems, which still limits their capability while handling the camera variations in vehicle Re-ID. Different from the previous Re-ID methods, our method jointly models multiple vehicle images guided by the camera topology graph in a unified framework.

B. Graph Convolutional Network (GCN)

In recent years, Graph Convolutional Network (GCN) [11] has become popular. GCN generalizes the capability of Convolutional Neural Network (CNN) by performing convolution operations on graph-structured data. Traditional GCN models are widely used in computer vision tasks such as pose estimation [29], [30], action recognition [31], [32], person re-identification [33], [34], vehicle re-identification [35], [6], etc. Zhao et al. [29] learn to capture pose information such as local and global node relationships by semantic graph convolutional

network (SemGCN). Hu et al. [30] propose spatio-temporal conditional directed graph convolution to leverage varying non-local dependence for different poses. Li et al. [31] propose actional-structural graph convolution network (AS-GCN) to extract useful spatial and temporal information for action recognition. Chen et al. [32] propose multi-scale temporal graph convolution to enrich the receptive field of the model in spatial and temporal dimensions for action recognition. For person re-identification, Shen et al. [33] propose a similarity-guided graph neural network to incorporate the rich gallery-gallery similarity information into the training process. Zhang et al. [34] propose a heterogeneous local graph attention network to model the inter-local relation and the intra-local relation in a unified framework for person Re-ID. However, these relations originate from sample similarity and are disturbed by visual representation models. For vehicle re-identification, Liu et al. [35] propose a parsing-guided cross-part reasoning network (PCRNet) to learn discriminative feature representations and model the correlation among parts. Shen et al. [6] propose a hybrid pyramidal graph network (HPGN) to explore the spatial significance of feature tensors at multiple scales for vehicle re-identification. In contrast, our proposed method considers potential topological relations under realistic traffic scenes. More specifically, this paper proposes a camera topology-based graph convolutional network to build a bridge between visual representation models and Closed Circuit Television camera systems.

III. Method

To connect Closed Circuit Television (CCTV) camera systems and vehicle representation models, we propose a camera topology graph guided vehicle re-identification framework, as shown in Fig. 3. We first describe the vehicle Re-ID setting and the vanilla GCN [11] settings for learning topological feature, followed by the Camera Topology Graph Construction and introduce the Camera Topology-based Graph Convolutional Network to propagate messages and update node features. Finally, we describe the Topological Cross-entropy Loss for the whole network training.

A. Vehicle Re-ID Setting

Vehicle re-identification aims to retrieve vehicles of interest across non-overlapping cameras. Given the training set $T = \{\mathbf{x}_i, y_i, y_i^{cam}\}_{i=1}^{N^T}$, where $\mathbf{x}_i$ denotes the i-th image, N^T represents the number of images in the training set. $y_i \in S_T$ is the corresponding identity label, where S_T contains the identities of all the training vehicle images. $y_i^{cam} \in C_T$ is the corresponding camera label, where C_T contains the cameras in the training set. During the training phase, we learn a vehicle representation model $\mathbf{h}_i = F(\mathbf{x}_i)$ that extracts discriminative vehicle representations $\mathbf{h}_i \in \mathbb{R}^{d\times 1}$ per vehicle image. In the testing phase, we have a query set $U = \{\mathbf{x}_i\}_{i=1}^{N^U}$ with vehicles of interest, where N^U represents the number of images in the query set. Then given a gallery set $G = \{\mathbf{x}_i\}_{i=1}^{N^G}$ for retrieval, we retrieve correct vehicles when comparing query images against the images in the gallery set G, where N^G represents the number of images in the gallery set. The identities of the vehicles in the query set S_U are disjoint from the identities available during the training phase, *i.e.*, $S_U \bigcap S_T = \emptyset$. It is worth noting that $C_U \subseteq C_T$, which means that the training set must contain the cameras in the query set. $C_U \subseteq C_T$ is the basic condition of vehicle Re-ID methods. If $C_U \bigcap C_T = \emptyset$, then the problem of Re-ID translates to Unsupervised Domain Adaptation Re-ID, which is discussed in other related studies [36], [37], [38].

B. Learning Topological Features by vanilla GCN

In the vehicle Re-ID setting, we have no additional conditions on the vehicle representation function $F(\cdot)$. Typically, $F(\cdot)$ is computed on a single vehicle image, thus ignoring any possible topological relations that may arise between the representations of the same vehicle across cameras. To explicitly account for such topological relations, we introduce an aggregation function to update the image representation vector, which can be formulated as:

$$\mathbf{h}_i^{'} = agg(\mathbf{h}_i, \{\mathbf{h}_j\}_{j=1}^{N}) = \sum_j \mathbf{h}_j \mathbf{w}_{ij}, \tag{1}$$

where $\{\mathbf{h}_j\}_{j=1}^{N}$ contains the representation vectors learned by the representation function $F(\cdot)$ of all the input images $\{\mathbf{x}_i\}_{i=1}^{N}$. During the training phase, $\{\mathbf{x}_i\}_{i=1}^{N} \subset T$ is a batch sampled from the training set. While during the testing phase, $\{\mathbf{x}_i\}_{i=1}^{N} = U \bigcup G$ contains all the vehicle images from the query set and the gallery set. $\mathbf{w}_{ij}$ is a learnable weight between $\mathbf{h}_i$ and $\mathbf{h}_j$, where $\sum_j \mathbf{w}_{ij} = 1$. Recently, vanilla Graph Convolutional Network (GCN) [11] has shown to be particularly apt in modeling relations between elements in a set. Inspired by this, we apply the vanilla GCN to the vehicle Re-ID problem. Eq. (1) can be rewritten as:

$$\mathbf{h}_i^{'} = agg(\mathbf{h}_i, \{\mathbf{h}_j\}_{j=1}^{N}) = \sigma(\sum_j \mathbf{M}\mathbf{h}_j norm(\mathbf{A})_{ij}), \tag{2}$$

where $\sigma(\cdot)$ is the activation function, *i.e.*, $ReLU$, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is an affinity matrix that contains the feature affinities between any two pairs of input representation vectors $\mathbf{h}_i$, $\mathbf{h}_j$, $norm(\cdot)$ is a L_2-normalization function to turn the affinities into weights, $\mathbf{M} \in \mathbb{R}^{d\times d}$ is a learnable transformation matrix.

Remark. To apply vanilla GCN model to the vehicle re-identification task, we believe that it is worth improving three aspects.

- First, the affinity matrix is obtained by calculating the visual feature similarities, ignoring the connection between vehicle representation models and CCTV camera systems.
- Second, the computational cost of feature aggregation is expensive, which presents a significant computational burden because of the large input size N.
- Third, the result of feature aggregation is not robust because it includes visual features and camera noise.

In the following subsections, we will describe the Camera Topology Graph Construction and introduce a novel Camera Topology-based Graph Convolutional Network (CT-GCN).

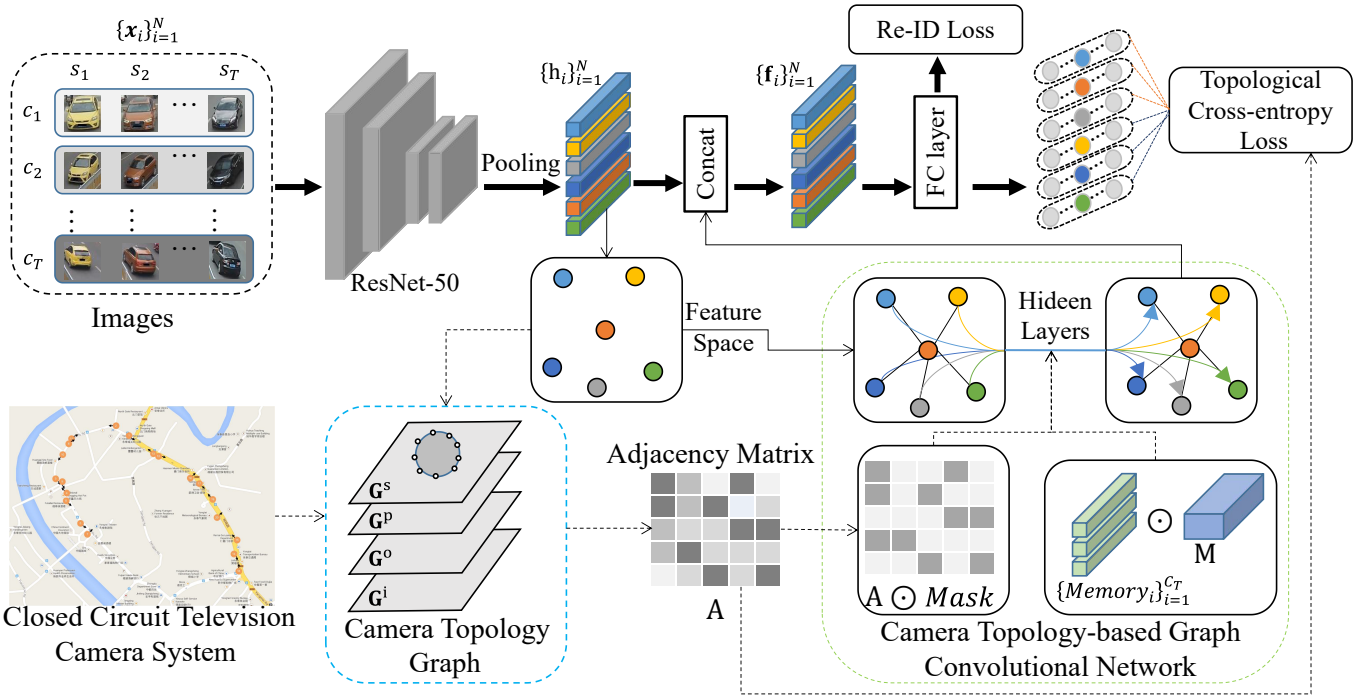


Fig. 3. Pipeline of Camera Topology Graph Guided Vehicle Re-identification framework. Given the images $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, we first extract the corresponding visual features $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\} \in \mathbb{R}^{d \times N}$ via ResNet-50. Next, we use the CCTV camera system guided camera topology graph to build an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ between visual features. Then, we dot-multiply adjacency matrix $\mathbf{A}$ by $Mask \in \mathbb{R}^{N \times N}$ to eliminate irrelevant visual features. Meanwhile, we weight transformation matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ by camera memory matrix $Memory \in \mathbb{R}^{d \times C_T}$ to store transformation matrices for different cameras. Visual features are transformed into topological features through adjacency relations and specific transformation matrices in Camera Topology-based Graph Convolutional Network. Moreover, we concatenate visual features and topological features into final vehicle features $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N\} \in \mathbb{R}^{d \times N}$. In addition to the commonly used Re-ID loss, we encourage positive samples to be clustered from strong to weak according to the topological relation in topological cross-entropy loss. Finally, our model is trained by the sum of Re-ID loss and topological cross-entropy loss.

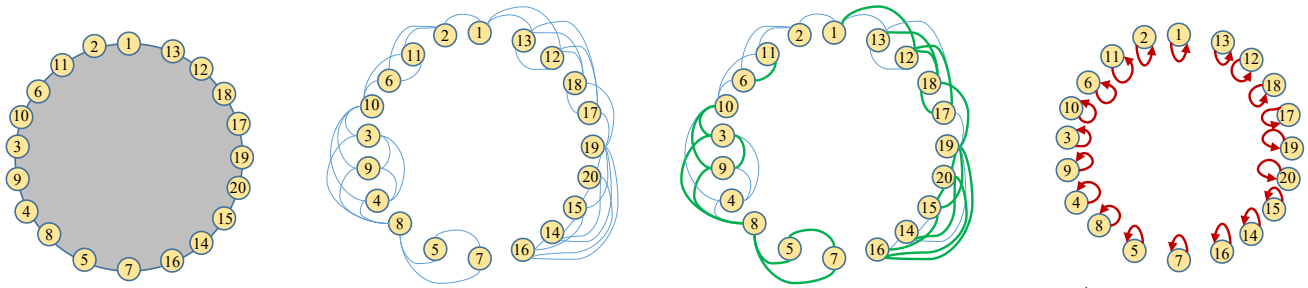
C. Camera Topology Graph Construction

To display the CCTV camera system in the form of graphs, we propose a simple and general method for constructing the Camera Topology Graph. The Camera Topology Graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ defines camera topological relations among different cameras. In fact, $\mathbf{V}$ represents camera nodes $\mathbf{V} = \{\mathbf{V}_1, ..., \mathbf{V}_{C_T}\}$, where C_T contains the camera numbers in the training set. $\mathbf{E}$ refers to the edge set in the camera topology graph, which contains four types of edges, *i.e.*, $\mathbf{E} = \{\mathbf{E}^s, \mathbf{E}^p, \mathbf{E}^o, \mathbf{E}^i\}$ denotes the four camera relationships based on camera system, position, orientation and individual, respectively. For clearer description, we represent the camera topology graph based on different relationships as $\{\mathbf{G}^s, \mathbf{G}^p, \mathbf{G}^o, \mathbf{G}^i\}$. Next, we will introduce these four camera relationships in detail.

(1) $\mathbf{G}^s$: Camera Topology Graph based on Camera System. The IDs of the positive samples are equivalent under the entire camera system. As shown in Fig. 4 (a), numbers $1-20$ represent the 20 cameras in the VeRi-776 dataset, the complete graph means that these 20 nodes have edges between each other. The camera relationship of $\mathbf{G}^s$ is the hardest because it requires positive samples from the entire camera system to present a consistent feature representation. Meanwhile, it is also the default relationship for all current vehicle re-identification methods.

(2) $\mathbf{G}^p$: Camera Topology Graph based on Camera Position. The closer the cameras are to each other, the more possible they are to capture positive samples. To clarify this graph relationship, we first define the cameras at two consecutive junctions as the nodes that are spatially adjacent. According to the camera position in the CCTV camera system (Fig. 2 (b)), we regard camera5, camera7, and camera8 as neighbor nodes, and there are edges between these neighbor nodes as shown in Fig. 4 (b). Compared with the camera relationship of $\mathbf{G}^s$, the camera relationship of $\mathbf{G}^p$ is easier because it requires positive samples from the neighbor cameras to present a consistent feature representation. Due to the continuously moving vehicle can be captured by two neighbor cameras, the camera relationship of $\mathbf{G}^{position}$ conforms to the vehicle driving logic. $\mathbf{G}^p$ aims to interact with positive samples under neighbor cameras.

(3) $\mathbf{G}^o$: Camera Topology Graph based on Camera Orientation. The more consistent the camera orientation, the more consistent the appearance from positive samples. As shown in Fig. 4 (c), camera3 and camera4 are neighbor cameras, there is no edge between them because their cameras are not oriented in the same direction. Compared with the camera relationship of $\mathbf{G}^p$, the camera relationship of $\mathbf{G}^o$ is easier because it ignores irrelevant nodes based on camera orientation. It is worth noting that we consider cameras whose two directions are orthogonal also as neighbor cameras, such



(a) $\mathbf{G}^{s}$: Camera Topology Graph based on Camera System

(b) $\mathbf{G}^{p}$: Camera Topology Graph based on Camera Position

(c) $\mathbf{G}^{o}$: Camera Topology Graph based on Camera Orientation

(d) $\mathbf{G}^{i}$: Camera Topology Graph based on Camera Individual

Fig. 4. Illustration of the proposed camera topology graph on VeRi-776 dataset [7]. which contains four camera relationships: (a) Camera System, (b) Camera Position, (c) Camera Orientation, (d) Camera Individual.

as camera5 and camera7 in Fig. 4 (c). $\mathbf{G}^{o}$ aims to interact with positive samples under cameras of consistent orientation.

(4) $\mathbf{G}^{i}$: Camera Topology Graph based on Camera Individual. A video sequence of the target vehicle can be captured under the same camera. As shown in Fig. 4 (d), any camera will have an edge over itself. The camera relationship of $\mathbf{G}^{i}$ is easiest because intra-class images captured under one camera tend to have a large information overlap. $\mathbf{G}^{i}$ aims to interact with positive samples under the same camera.

Learning camera system, position, orientation and individual relationships help to reduce the feature interaction range in the feature learning stage and the evaluation stage. In the above four subgraphs, if there is an edge between nodes, the value is 1, otherwise the value is 0. We use these four subgraphs together to build the camera topology graph. In a camera topology graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, the edges of two cameras can be expressed as $\mathbf{E}_{ij}$, and the larger the value, the stronger the relationship between the cameras. With four types of topological relations between cameras, our goal is to obtain hierarchically aggregated topological features. Such topological features are complementary to visual features, which make the final features more adequate and robust.

D. Camera Topology-based Graph Convolutional Network

After obtaining the camera topology graph, we design a Camera Topology-based Graph Convolutional Network (CT-GCN) to effectively propagate messages and update node features. To be specific, Given a batch of training data $T = \{\mathbf{x}_i, y_i, y_i^{cam}\}_{i=1}^{N}$, N represents batch size. The corresponding vehicle feature vector encoded by the network is denoted as $\mathbf{h}_i = F(\mathbf{x}_i)$. To embed the topological relationship into the feature representation model, we transform the topological relationship between cameras to the sample pairs. Formally, the topological relationship ($\mathbf{A}_{ij}$) of two vehicle feature vectors ($\mathbf{h}_i$ and $\mathbf{h}_j$) in the training set can be expressed as:

$$\mathbf{A}_{ij} = \mathbf{E}_{y_i^{cam} y_j^{cam}} \tag{3}$$

where $\mathbf{E}_{y_i^{cam} y_j^{cam}}$ denotes the edge between the y_i^{cam}-th camera and the y_j^{cam}-th camera in camera topology graph $\mathbf{G}$. It can be seen from Eq. (3) that we use the camera relationship between samples to represent the feature relationship between samples. This is because the stronger the camera relationship between the samples, the more overlap between the vehicle images as described in Section. III-C. However, this process incorporates many irrelevant samples and imposes a huge computational burden.

To discard irrelevant samples and reduce the computational burden, we introduce a mask matrix. We assume that if two vehicle images are visual neighbors in feature space, they are likely to be relevant. To this end, we propose to compute a top-k neighbor mask $Mask \in \mathbb{R}^{N \times N}$ from visual similarities, which will attend to the top-k value of similarities per row:

$$Mask_{ij} = \begin{cases} 1, & if\ j \in topk(Sim_{i,:}), \\ 0, & otherwise, \end{cases} \tag{4}$$

where $Sim_{i,j}$ denotes the feature similarity between the i-th image and the j-th image. For each element $Mask_{ij}$, the value will be set to 1 if j is top-k visual neighbor of i, 0 otherwise. By adding this mask matrix $Mask$ to the affinity matrix, we achieve feature aggregation only occurring in neighbors, which increases the focus on more relevant images. Compared with Eq. (2), the feature aggregation can be defined as follows:

$$\mathbf{h}_i^{'} = \sigma(\sum_j \mathbf{M}\mathbf{h}_j norm(Mask \odot \mathbf{A})_{ij}), \tag{5}$$

where $\odot$ is the element-wise product. Since most irrelevant visual neighbors are set to zero and thus relations are restricted to relevant neighbors, which makes the aggregation process more robust.

In addition, although Eq. (5) obtains more robust aggregated features while reducing computational complexity, this aggregation process introduces unwanted camera noise. The reason is that it shares a feature transformation matrix $\mathbf{M}$ for each node and thus prevents the graph convolutional network from learning camera-independent features. To solve this problem, we design a learnable camera memory matrix $Memory \in \mathbb{R}^{C_T \times d}$, where C_T contains the cameras in the training set. The motivation of the camera memory matric $Memory$ is to relax the original feature transformation matrix $\mathbf{M}$. Camera Topology-based Graph Convolutional Network uses a shared transformation matrix $\mathbf{M}$ as in the vanilla GCN

but learns a different camera memory vector $Memory_{y_i^{cam}}$ for each node i. Mathematically,

$$\mathbf{h}_i^{'} = \sigma(\sum_j (Memory_{y_i^{cam}} \odot \mathbf{M})\mathbf{h}_j norm(Mask \odot \mathbf{A})_{ij}), \tag{6}$$

where $Memory_{y_i^{cam}} \in \mathbb{R}^d$ is a learnable weight vector for camera y_i^{cam} ($y_i^{cam} = \{1, ..., C_T\}$), $\odot$ denotes element-wise multiplication but should broadcast properly. Specifically, $(Memory_{y_i^{cam}} \odot \mathbf{M})$ means the d-th row of $\mathbf{M} \in \mathbb{R}^{d \times d}$ is scaled by the d-th element of $Memory_{y_i^{cam}}$, and the result is of the same dimension as $\mathbf{M}$, *i.e.*, $d \times d$. $\mathbf{M}$ converts nodes from different cameras equally, $(Memory_{y_i^{cam}} \odot \mathbf{M})$ represents the feature transformation matrix for y_i^{cam}-th camera. From Eq. (6), we can understand the camera memory matrix as storing memory items to fine-tune the feature transformation matrix, which obtains camera-independent features and avoids aggregating camera noise during the feature aggregation process.

E. Topological Cross-entropy Loss

After obtaining the topological feature $\mathbf{h}_i^{'}$, it is concatenated with the visual feature $\mathbf{h}_i$ and denoted as $\mathbf{f}_i = Concat(\mathbf{h}_i, \mathbf{h}_i^{'})$. Note that $\mathbf{h}_i$ is the original visual feature, and $\mathbf{h}_i^{'}$ is the combination of neighbor features. The joint representation $\mathbf{f}_i$ learns more meaningful features by the network. Then $\mathbf{f}_i$ is fed into a fully-connected layer to further obtain class prediction results. The class prediction results are finally optimized by Re-ID loss $\mathcal{L}_{ReID}$ in the form of,

$$\begin{aligned} \mathcal{L}_{ReID} = &-y_i log(Softmax(FC(\mathbf{f}_i))) + \\ &max(0, \|\mathbf{f}_i - \mathbf{f}_{i,p})\| + m - \|\mathbf{f}_i - \mathbf{f}_{i,n})\|), \end{aligned} \tag{7}$$

where FC denotes a fully-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 i, p and i, n indicate the hardest positive and hardest negative feature index in each mini-batch for the sample $\boldsymbol{x}_i$, and $m = 0.3$ denotes the triplet distance margin. $\mathcal{L}_{ReID}$ denotes the widely-used cross-entropy loss [39], and triplet loss [40] with batch hard mining on the Re-ID feature vectors. Although these two loss functions derived from image classification tasks are widely used in the field of vehicle re-identification, they have a limitation in that they can not consider the topological relationship between samples.

As described in Section. III-C, the stronger the camera relationship between the samples, the more overlap between the vehicle images. Our key idea is to preferentially train the easy-positive samples instead of training all positive samples at once. To incorporate the camera topological relationship in the training phase, we first calculate the class prediction results of the anchor sample and neighbor samples. We then give different learning weights of neighbor samples based on camera topological relationship. Finally, we propose a novel topological cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{TCE} = -\frac{1}{S_i} \sum_{j=1}^{S_i} &norm(\mathbf{A})_{ij} norm(Softplus(FC(\mathbf{f}_j))) * \\ &log(Softmax(FC(\mathbf{f}_i))), \end{aligned} \tag{8}$$

where S_i represents the number of positive samples of the $i-$th image, $norm(\mathbf{A})_{ij}$ denotes the topological relationship between two positive pairs $\mathbf{f}_i$ and $\mathbf{f}_j$. $Softplus$ denotes the Softplus function that gets the non-negative probability. The topological cross-entropy loss forces positive samples to learn vehicle representation from easy to hard. The final objective function for our model rewrite as:

$$\mathcal{L}_{total} = \mathcal{L}_{ReID} + \lambda \mathcal{L}_{TCE}, \tag{9}$$

where only λ is used to balance the Re-ID loss and topological cross-entropy loss.

In the inference/testing stage, we first use the pre-trained ResNet-50 [17] to learn visual features. Then we send the visual features and the adjacent matrix to the pre-trained CT-GCN to learn topological features. Eventually we connect visual features and topological features to obtain the final features for identification. It is worth noting that camera labels of the testing images are used to build adjacent matrix and filter out samples from the gallery with the same ID and the same camera as the query.

IV. Experiments

We evaluate our method on three vehicle Re-ID datasets VeRi-776 [7], VERI-Wild [41] and VehicleID [18] comparing to the state-of-the-art methods in this section. We adopt the mean average precision (mAP), Rank-1 (R-1) accuracy and Rank-5 (R-5) accuracy as the evaluation metric. mAP measures the mean of all queries of average precision (the area under the Precision Recall curve) which reflects the recall. Rank-score is an estimation of finding the correct match in the Top-k returned results.

A. Datasets

VeRi-776 dataset [7] consists of 49357 images of 776 distinct vehicles captured by 20 non-overlapping cameras in different orientations and lighting conditions. Among them, 576 identities (37778 images) and 200 identities (11579 images) are assigned as training and testing respectively. Furthermore, 1678 images from 200 identities are selected as the query from the testing set.

VERI-Wild dataset [41] is a large-scale dataset containing 416314 images of 40671 vehicles captured by 174 cameras. The training set consists of 277797 images of 30671 vehicles. There are 138517 images of 10000 identities in the test set, which consists of three different scale testing subsets, *i*.e., Test3000 (Small), Test5000 (Medium), and Test10000 (Large). In this dataset, we do not know the specific position and orientation of the camera and thus are unable to construct a camera topology graph directly. Instead, we use camera labels to train a camera classification network and then use it to calculate the features and relationships of cameras. In this way, we can also achieve the purpose of constructing the camera topology graph.

VehicleID dataset [18] is a large-scale dataset used for vehicle retrieval tasks and is composed of 221567 images from 26328 unique vehicles. The training set contains 110178 images of 13134 vehicles, while the testing set contains 111585 images

TABLE I
COMPARISON RESULTS OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VERI-776 DATASET [7].

Methods	mAP	Rank-1	Rank-5
LOMO [42]	0.096	0.253	0.465
GoogLeNet [43]	0.170	0.498	0.712
FACT [44]	0.188	0.522	0.729
SCPL [8]	0.583	0.835	0.900
OIFE [9]	0.480	0.659	0.877
NuFACT [45]	0.485	0.769	0.914
VAMI [1]	0.501	0.770	0.908
GSTE [20]	0.578	0.958	0.965
EALN [2]	0.574	0.844	0.941
AAVER [4]	0.612	0.890	0.947
VANet [3]	0.663	0.898	0.960
PRN [21]	0.743	0.943	0.989
UMTS [46]	0.759	0.958	-
PVEN [5]	0.795	0.956	0.984
SAVER [24]	0.796	0.964	0.986
HPGN [6]	0.802	0.967	-
MsKAT [23]	0.820	0.971	0.990
Baseline [10]	0.766	0.957	0.980
OURS	**0.827**	**0.971**	**0.990**

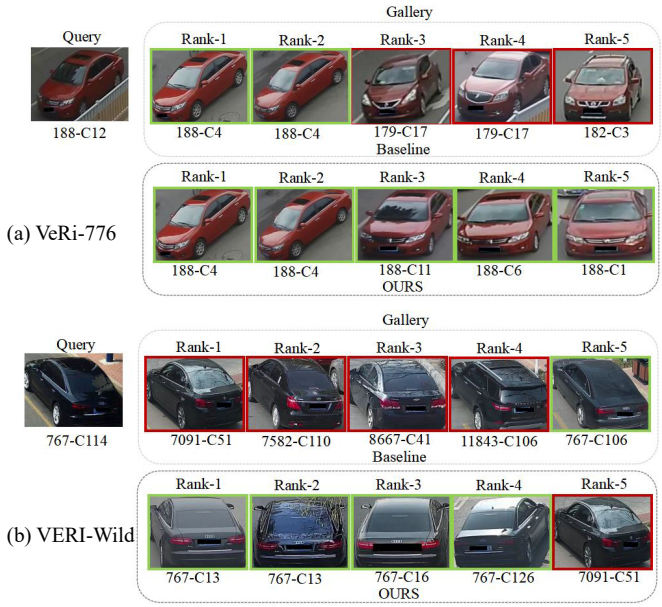


Fig. 5. Top five returned results of the baseline model versus our model on VeRi-776 and VERI-Wild datasets. 188-C12 means that the identity label of the vehicle is 188 and the camera label is 12. The images with green bounding boxes and the rest ones indicate the correct and wrong matchings respectively.

of 13133 vehicles. There are 6 testing splits with various gallery sizes as 800, 1600, 2400, 3200, 6000, and 13164. Following the protocol in [18], [21], [5], we use the first three splits Small (S), Medium (M) and Large (L) for testing. During the testing phase, one single image of each identity is randomly selected to form the gallery set while the rest of the images are as the query. This procedure is repeated ten times and the averaged metrics. Furthermore, the camera position, orientation and label are also not available. To better evaluate our proposed model, we generate a camera topology graph by the pre-trained camera classification network from VERI-Wild dataset [41] for VehicleID dataset [18].

B. Implementation details

In our experiments, we adopt ResNet-50 [17] pretrained on ImageNet [48] without the last spatial down-sampling layer as the backbone model followed by [10]. We use the Adam [49] optimizer with the initial learning rate of $3.5e-5$. We adopt a warmup [50] mechanism to bootstrap the network, which takes 10 epochs to linearly increase the learning rate from $3.5e-5$ to $3.5e-4$. The learning rate decays to $3.5e-5$ and $3.5e-6$ at the 40-th epoch and the 70-th epoch respectively (overall 120 epochs). The training protocol follows the Re-ID strong baseline (BOT [10]) using random cropping and erasing for data augmentation. In our implementation, all the input images are resized to 256×256. The dimension of both visual and topological features is $d = 2048$. We set 16 IDs, and 4 instances with the batch size of 64 in the training for the three datasets. We run our experiments on two NVIDIA GeForce RTX 2080Ti GPUs with 11GB RAM. Compared with the baseline model, we add two parameter tensors that need to be trained, which are the camera memory matrix $Memory$ and the feature transformation matrix M. The $Memory$ weights and M weights are randomly initialized.

TABLE IV
COMPARISON RESULTS OF OUR METHOD AGAINST THE STATE-OF-THE-ART METHODS ON VEHICLEID DATASET [18].

Methods	Small		Medium		Large	
	R-1	R-5	R-1	R-5	R-1	R-5
LOMO [42]	0.197	0.321	0.189	0.295	0.153	0.256
GoogLeNet [43]	0.479	0.674	0.435	0.635	0.382	0.595
FACT [44]	0.495	0.680	0.446	0.642	0.399	0.605
OIFE [9]	-	-	-	-	0.670	0.823
VAMI [1]	0.631	0.833	0.529	0.751	0.473	0.703
RAM [51]	0.752	0.915	0.723	0.870	0.677	0.845
EALN [2]	0.751	0.881	0.718	0.839	0.693	0.814
AAVER [4]	0.747	0.938	0.686	0.900	0.635	0.856
VANet [3]	0.833	0.960	0.811	0.947	0.772	0.929
PRN [21]	0.784	0.923	0.750	0.883	0.742	0.864
PVEN [5]	0.847	0.970	0.806	0.945	0.778	0.920
HPGN [6]	0.839	-	0.800	-	0.773	-
MsKAT [23]	0.863	0.974	0.818	0.955	0.794	0.939
Baseline [10]	0.802	0.914	0.775	0.887	0.738	0.849
OURS	**0.870**	**0.979**	**0.826**	**0.962**	**0.806**	**0.944**

C. Comparison to State-of-the-art Methods

Evaluation Results on VeRi-776 [7]. Table I reports the performance comparison of our method against the state-of-the-art methods on VeRi-776 dataset [7]. We can observe that our approach significantly beats the state-of-the-art methods as 82.7% on mAP. Although the second-best method MsKAT [23] outperforms the third-best one HPGN [6] by a large margin by exploring the identity-related information in the vehicle image, it requires additional attribute and state annotations. HPGN [6] learns hierarchical part features through traditional graph convolutional networks. However, HPGN [6] lacks consideration of removing camera noise and only interacts with the part features. By capturing the topological information across cameras in different images, our method learns more robust feature representations as

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

Methods	Small			Middle			Large		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Unlabled-GAN [47]	0.299	0.581	0.796	0.247	0.516	0.744	0.182	0.436	0.655
GSTE [20]	0.314	0.605	0.801	0.262	0.521	0.749	0.195	0.454	0.665
FDA-Net [41]	0.351	0.640	0.828	0.298	0.578	0.783	0.228	0.494	0.705
UMTS [46]	0.727	0.845	-	0.661	0.793	-	0.542	0.728	-
AAVER [4]	0.622	0.758	0.927	0.537	0.682	0.889	0.417	0.587	0.876
SAVER [24]	0.809	0.945	0.981	0.753	0.927	0.974	0.677	0.895	0.958
PVEN [5]	0.825	0.967	0.992	0.770	0.954	0.988	0.697	0.934	0.978
HPGN [6]	0.804	0.914	-	0.752	0.882	-	0.650	0.827	-
MsKAT [23]	0.840	0.973	0.993	0.787	0.956	0.990	0.722	0.939	0.983
Baseline [10]	0.762	0.918	0.966	0.680	0.873	0.945	0.578	0.835	0.917
OURS	**0.860**	**0.973**	**0.996**	**0.812**	**0.958**	**0.991**	**0.734**	**0.939**	**0.985**

TABLE III
ABLATION STUDY ON VERI-776, VEHICLEID AND VERI-WILD.

Variant	VeRi-776		VehicleID						VERI-Wild					
			Small		Medium		Large		Small		Medium		Large	
	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) Baseline	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) w/o CTG	0.798	0.961	0.836	0.957	0.803	0.946	0.764	0.914	0.836	0.958	0.776	0.924	0.669	0.906
(c) w/o CT-GCN	0.792	0.963	0.821	0.938	0.794	0.925	0.753	0.871	0.794	0.940	0.744	0.912	0.638	0.884
(d) w/o TCE Loss	0.815	0.968	0.856	0.970	0.810	0.958	0.780	0.939	0.853	0.969	0.803	0.946	0.726	0.930
(e) **OURS**	**0.827**	**0.971**	**0.870**	**0.979**	**0.826**	**0.962**	**0.806**	**0.944**	**0.860**	**0.973**	**0.812**	**0.958**	**0.734**	**0.939**

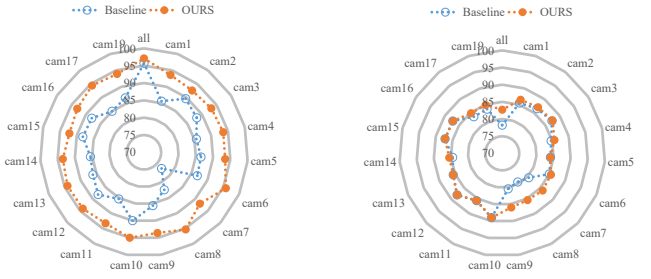
shown in Fig. 5 (a).

Evaluation Results on VehicleID [18]. Table IV shows the comparison results of VehicleID [18] on three different testing sets. We compare the Rank-1 (R-1) and Rank-5 (R-5) scores on this dataset since there is only one ground-truth for each query in the gallery. Our method merges the query and gallery into a collection that requires feature interaction. In general, our method achieves promising performance compared to state-of-the-art methods. The main reason is, that previous methods consider learning representations from a single vehicle image, ignoring any potential interactions between images. Compared with the baseline, our proposed model significantly improves Rank-1 by 6.8%, 5.1%, and 6.8% on three different testing sets respectively. This shows the promising achievement of using Graph Convolutional Network for feature aggregation under multiple neighbor images.

Evaluation Results on VERI-Wild [41]. Table II reports the comparison results of VERI-Wild [41] on three different testing sets. As shown in Table II, our approach significantly beats the second-best method MsKAT [23] by 86.0%, 81.2% and 73.4% on three different testing sets respectively. Through the effective interaction of feature vectors and graph representation learning, our proposed approach improves the mAP of three different testing sets by 2.0%, 2.5%, and 1.2% respectively. As shown in Fig. 5 (b), it can be observed that our method outperforms the retrieval performance of baseline model, which indicates the robust generalization ability of the proposed model in large-scale datasets.

D. Ablation Study

The contribution components of our model are mainly in three aspects, Camera Topology Graph (CTG), Camera Topology-based Graph Convolutional Network (CT-GCN), and Topological Cross-entropy Loss (TCE Loss).



(1) Rank-1 performance under different cameras (2) mAP performance under different cameras

Fig. 6. Comparison results of our model against the Baseline model under different cameras on VeRi-776 dataset.

TABLE V
ANALYSIS OF CAMERA TOPOLOGY GRAPH ON VERI-776 (IN %).

	Component	mAP	Rank-1	Rank-5
1)	Baseline	76.6	95.7	98.0
2)	+ $\mathbf{G}^s$	79.8	96.1	98.5
3)	+ $\mathbf{G}^p$	80.5	96.7	98.7
4)	+ $\mathbf{G}^o$	80.2	96.5	98.7
5)	+ $\mathbf{G}^i$	77.5	95.8	98.4
6)	+ $\mathbf{G}^s$ + $\mathbf{G}^p$	80.9	96.8	98.7
7)	+ $\mathbf{G}^s$ + $\mathbf{G}^p$ + $\mathbf{G}^o$	82.3	96.9	98.9
8)	+ $\mathbf{G}^s$ + $\mathbf{G}^p$ + $\mathbf{G}^o$ + $\mathbf{G}^i$	82.7	97.1	99.0

Effectiveness of Each Component. To verify the contribution of the components in our model, we implement several variants of our method on the three datasets, as reported in Table III. Our baseline model is ResNet-50 pretrained on ImageNet [48], which follows the Re-ID strong baseline [10] in the experimental setting. Since the Camera Topology Graph (CTG) is an adjacency matrix, which can only be used in combination with the Camera Topology-based Graph Convolutional Network (CT-GCN) or Topological Cross-entropy Loss (TCE Loss). In

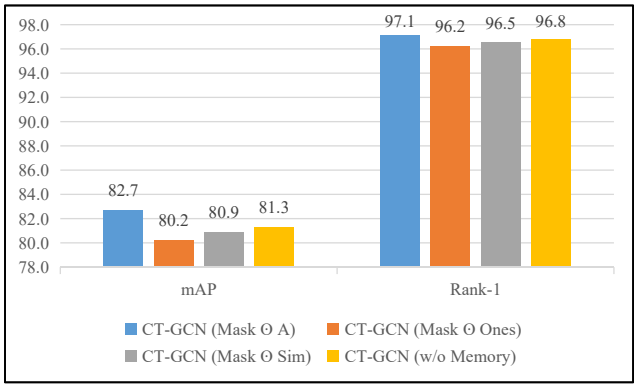


Fig. 7. Analysis of Camera Topology-based Graph Convolutional Network.

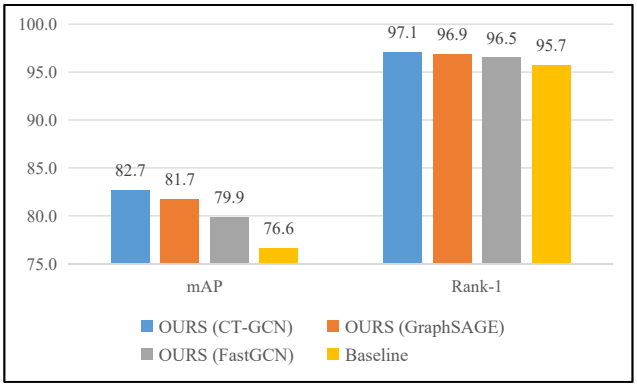


Fig. 8. Analysis of different GCN models on VeRi-776.

the first variant (OURS w/o CTG), we consider the complete graph of the camera-based system $\mathbf{G}^s$, and then keep the Camera Topology-based Graph Convolutional Network (CT-GCN) or Topological Cross-entropy Loss (TCE Loss).

Comparing Table III (b) and Table III (a), we can find that the Camera Topology Graph (CTG) plays an important role in the entire model, which indicates that topological relations between images can effectively learn cross-camera representations. By removing the Camera Topology-based Graph Convolutional Network (CT-GCN, Table III (c)), and the Topological Cross-entropy Loss (TCE Loss, Table III (d)) respectively, both mAP, and Rank-1 scores significantly decrease on all the three datasets with different test settings. This demonstrates the effectiveness of each component in our method. In addition to this, we compare the performance results of our model and the baseline model under different cameras. As shown in Fig. 6 (1), Our model significantly improves the Rank-1 performance under different cameras, which indicates that the proposed model can effectively mitigate the influence of camera changes on Rank-1 performance. As shown in Fig. 6 (2), our model mainly improves mAP performance under several more challenging cameras, which indicates that learning cross-camera feature representations can effectively alleviate the influence of complex cameras on mAP performance.

The limitation of the proposed framework can be seen in Fig. 6 (2), our method cannot improve mAP performance compared to the baseline model under each camera. The potential reason may be that the proposed method does not balance the importance of camera relationships. To address this issue, in the future, we will consider designing a multi-head graph convolutional network that allows us to dynamically assign different weights for different graph relationships.

Analysis of Camera Topology Graph. The Camera Topology Graph (CTG) is an adjacency matrix determined by the CCTV camera system. The graph $\mathbf{G}$ is actually a combination of four camera relationships $\mathbf{G} = \{\mathbf{G}^s, \mathbf{G}^p, \mathbf{G}^o, \mathbf{G}^i\}$ as described in Section. III-C. To better evaluate four camera relationships of our model, we consider several variants on VeRi-776 dataset by progressively introducing the $\mathbf{G}^s$, $\mathbf{G}^p$, $\mathbf{G}^o$ and $\mathbf{G}^i$ into the "Baseline" as shown in Table. V. Clearly, both mAP,

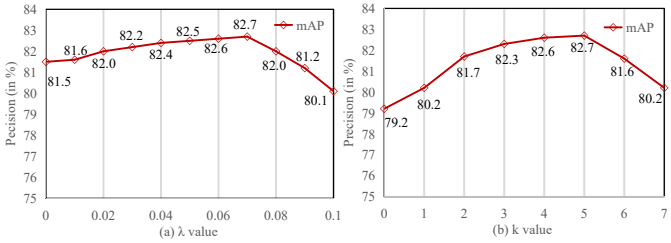


Fig. 10. Parameter analysis at mAP on VeRi-776 dataset.

and Rank-1 scores consistently increase, which verifies the effectiveness of each graph relation.

Analysis of Camera Topology-based Graph Convolutional Network. As we discussed, the Camera Topology-based Graph Convolutional Network (CT-GCN) captures the relationship between the vehicle images from different cameras, extracts and embeds topological features into the visual features for vehicle Re-ID. Compared with the traditional GCN, the proposed CT-GCN has two main differences: *i.e.*, the camera topology graph guided adjacency matrix A and the camera memory matrix $Memory$. To better analyze the proposed CT-GCN, we consider several variants of CT-GCN. 1) CT-GCN ($Mask \odot \mathbf{A}$) represents the proposed CT-GCN guided by the camera topology graph. 2) CT-GCN ($Mask \odot Ones$) denotes the value of the matrix is all 1. 3) CT-GCN ($Mask \odot Sim$) indicates that the value of the matrix is the feature similarity. 4) CT-GCN ($w/o\ Memory$) denotes the proposed CT-GCN without camera memory matrix $Memory$. As shown in Fig. 7, embedding the camera topology graph significantly increases all the metrics on VeRi-776 [7] dataset. We introduce the camera memory matrix $Memory$ into the CT-GCN. Both mAP, Rank-1 and Rank-5 scores significantly increase on VeRi-776 [7] dataset as shown in Fig. 7.

To further verify the effectiveness of the proposed CT-GCN, we use other GCN-based methods to replace the proposed CT-GCN. Specifically, we compare our method with GraphSAGE [54] and FastGCN [55], as reported in Fig. 8. It is worth noting that GraphSAGE [54] is similar to our method of sampling neighbors. Compared to FastGCN [55], GraphSAGE [54] has exhibited markedly superior performance, demonstrating the effectiveness of sampling neighbors

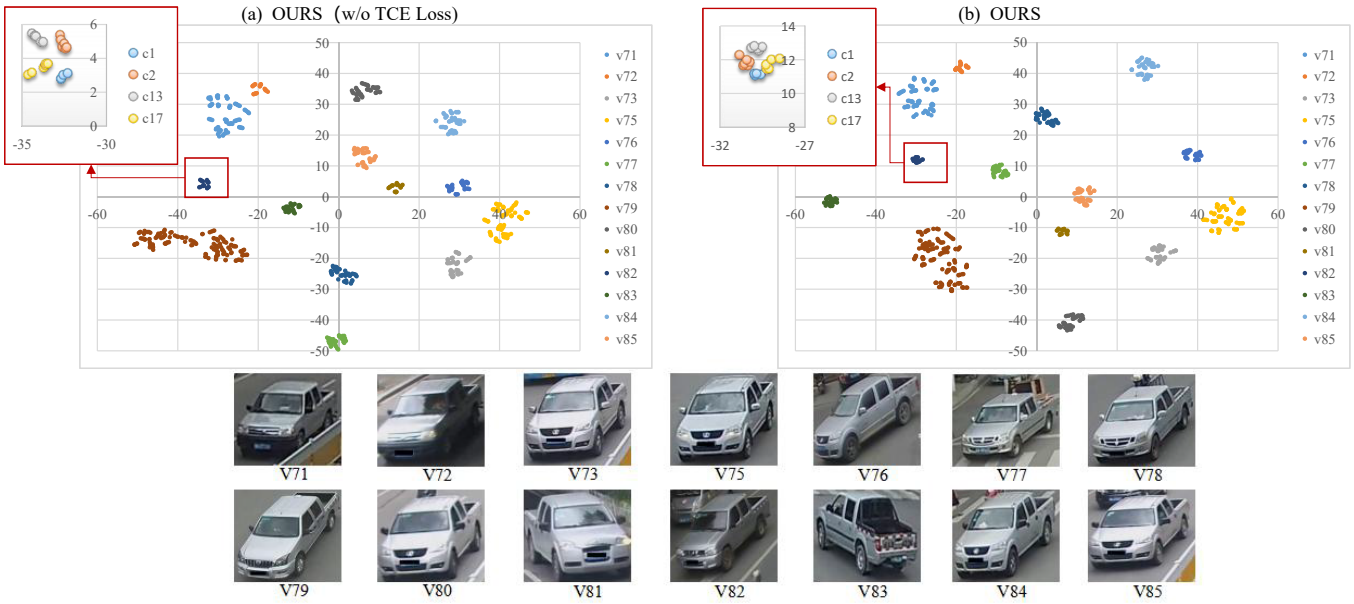


Fig. 9. Feature visualization of fourteen vehicle identities T-SNE [52]. Fourteen identities with gray color and pickup type are selected from the VeRi-776 and their IDs are listed on the right side of graphs. The nodes in different colors indicate the image features of different vehicles.

TABLE VI
COMPARISON RESULTS OF DIFFERENT BASELINE MODELS ON VERI-776, VEHICLEID AND VERI-WILD.

Variant	VeRi-776		VehicleID						VERI-Wild					
			Small		Medium		Large		Small		Medium		Large	
	mAP	R-1	R-1	R-5	R-1	R-5	R-1	R-5	mAP	R-1	mAP	R-1	mAP	R-1
Strong Baseline [10]	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
Strong Baseline [10] + OURS	**0.827**	**0.971**	**0.870**	**0.979**	**0.826**	**0.962**	**0.806**	**0.944**	**0.860**	**0.973**	**0.812**	**0.958**	**0.734**	**0.939**
FastReID [53]	0.804	0.965	0.823	0.955	0.807	0.927	0.778	0.901	0.819	0.963	0.757	0.945	0.667	0.911
FastReID [53] + OURS	**0.831**	**0.973**	**0.865**	**0.976**	**0.828**	**0.957**	**0.810**	**0.941**	**0.854**	**0.970**	**0.808**	**0.951**	**0.721**	**0.932**
HRCN [25]	0.818	0.964	0.873	0.980	0.817	0.961	0.797	0.943	0.842	0.940	0.791	0.927	0.710	0.899
HRCN [25] + OURS	**0.836**	**0.973**	**0.882**	**0.983**	**0.835**	**0.967**	**0.831**	**0.945**	**0.866**	**0.952**	**0.818**	**0.948**	**0.734**	**0.913**

for the vehicle Re-ID task. Compared to GraphSAGE [54], the proposed CT-GCN has achieved higher performance, the reason may be that it takes into account the camera memory matrix to extract various camera representation functions. This verifies that the Camera Topology-based Graph Convolutional Network (CT-GCN) guides more discriminative feature learning for vehicle Re-ID.

Analysis of Topological Cross-entropy Loss. The Topological Cross-entropy Loss aims to close the distance between positive samples by topological relations. To better visualize the contribution of the Topological Cross-entropy Loss, We show the visual feature distribution graph as shown in Fig. 9. From Fig. 9 (a), we can see that the model variant (OURS w/o TCE Loss) obtains suboptimal vehicle features at visualization space. Especially, the features of "V82" appear to be aggregated under the same camera, and still appear discretely distributed under different cameras ("c1", "c2", "c13", and "c17"). This in turn means the inter-class distance and the intra-class similarity are not well guaranteed in the model variant (OURS w/o TCE Loss). By considering the proposed Topological Cross-entropy Loss, our model significantly improves the feature learning by aggregating the image features of the same vehicle from different cameras, as shown in Fig. 9 (b). This shows that our method learns more robust feature representations against the diverse camera changes for the vehicle re-identification task.

E. Parameter Analysis

There are two important parameters in our model. λ balances the Re-ID Loss and Topological Cross-entropy Loss, while k control the number of neighbor nodes respectively. For the balanced parameter λ, as shown in Fig. 10 (a), we can observe that it is relatively insensitive when we slightly adjust it in range (0.02, 0.07). The value of the balanced parameter λ cannot be too large, as it will affect the training of the ground-truth labels. For the top-k of numbers of neighbors, as shown in Fig. 10 (b), we find that the performance of our method increases until $k = 5$. The number of neighbor nodes can not be too large, it will aggregate the information of irrelevant samples. As a result, we empirically set $\lambda = 0.07$, $k = 5$.

F. Baseline Analysis

To verify the effectiveness of our proposed model, we change the baseline model of our method and implement the baseline study on VeRi-776, VehicleID and VERI-Wild datasets, as reported in Table VI. Specifically, we adopt Strong Baseline [10], FastReID [53], HRCN [25] as our baseline

models. From Table VI we find that all three baseline models have improved performance by introducing our method, which demonstrates the effectiveness of our proposed model. It is important to note that these baseline models focus on the acquisition of visual information, while our method focuses on the supplementation of visual information. In other words, our method can be seen as a plug-and-play module that is applicable to integrate into any vehicle representation model.

V. Conclusion

We propose a novel camera topology graph guided vehicle re-identification framework for topological vehicle feature aggregation in end-to-end learning. Specifically, we propose a Camera Topology Graph to build a connecting bridge between vehicle representation models and CCTV camera systems under real-world traffic scenarios. Moreover, we design a novel Camera Topology-based Graph Convolutional Network and Topological Cross-entropy loss to learn more robust cross-camera features for the Re-ID task. Our method achieves superior performance and offers a more reasonable solution for vehicle Re-ID regarding cross-camera recognition. It is a trend for vehicle re-identification tasks to closely link vehicle representation models with real-world traffic systems. In the future, we will combine more information in the form of a multi-head graph convolutional network to establish the more powerful vehicle Re-ID solution.

References

[1] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6489–6498.

[2] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3794–3807, 2019.

[3] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8282–8291.

[4] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J. Chen, and R. Chellappa, "A dual path model with adaptive attention for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6132–6141.

[5] D. Meng, L. Li, X. Liu, Y. Li, S. Yang, Z.-J. Zha, X. Gao, S. Wang, and Q. Huang, "Parsing-based view-aware embedding network for vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7103–7112.

[6] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[7] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 869–884.

[8] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1918–1927.

[9] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 379–387.

[10] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, pp. 1–12, 2019.

[11] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the International Conference on Learning Representations*, 2017.

[12] K. Lv, H. Du, Y. Hou, W. Deng, H. Sheng, J. Jiao, and L. Zheng, "Vehicle re-identification with location and time stamps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 399–406.

[13] M. V. Prasad, R. Balakrishnan *et al.*, "Spatio-temporal association rule based deep annotation-free clustering (star-dac) for unsupervised person re-identification," *Pattern Recognition*, vol. 122, p. 108287, 2022.

[14] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157–5166.

[15] S. Zhou, M. Ke, and P. Luo, "Multi-camera transfer gan for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 393–400, 2019.

[16] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176–1190, 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[18] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.

[19] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 562–570.

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

[21] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3997–4005.

[22] X. Liu, S. Zhang, X. Wang, R. Hong, and Q. Tian, "Group-group loss-based global-regional feature learning for vehicle re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 2638–2652, 2019.

[23] H. Li, C. Li, A. Zheng, J. Tang, and B. Luo, "Mskat: Multi-scale knowledge-aware transformer for vehicle re-identification," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2022.

[24] P. Khorramshahi, N. Peri, J. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 369–386.

[25] J. Zhao, Y. Zhao, J. Li, K. Yan, and Y. Tian, "Heterogeneous relational complement for vehicle re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 205–214.

[26] H. Li, C. Li, A. Zheng, J. Tang, and B. Luo, "Attribute and state guided structural embedding network for vehicle re-identification," *IEEE Transactions on Image Processing*, pp. 1–14, 2022.

[27] J. Sochor, A. Herout, and J. Havel, "Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3006–3015.

[28] Y. Zhou and L. Shao, "Cross-view gan based vehicle generation for re-identification," in *Proceedings of the British Machine Vision Conference*, vol. 1, 2017, pp. 1–12.

[29] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3425–3435.

[30] W. Hu, C. Zhang, F. Zhan, L. Zhang, and T.-T. Wong, "Conditional directed graph convolution for 3d human pose estimation," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 602–611.

[31] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3595–3603.

[32] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1113–1122.

[33] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 486–504.

[34] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 136–12 145.

[35] 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 *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 907–915.

[36] F. Yang, Z. Zhong, Z. Luo, Y. Cai, Y. Lin, S. Li, and N. Sebe, "Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4855–4864.

[37] T. Isobe, D. Li, L. Tian, W. Chen, Y. Shan, and S. Wang, "Towards discriminative representation learning for unsupervised person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 8526–8536.

[38] D. Cheng, J. Zhou, N. Wang, and X. Gao, "Hybrid dynamic contrast and probability distillation for unsupervised person re-id," *IEEE Transactions on Image Processing*, vol. 31, pp. 3334–3346, 2022.

[39] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1735–1742.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[41] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3235–3243.

[42] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206.

[43] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3973–3981.

[44] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proceedings of the International Conference on Multimedia and Expo*, 2016, pp. 1–6.

[45] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2018.

[46] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 165–11 172.

[47] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

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

[50] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 51–58, 2019.

[51] X. Liu, S. Zhang, Q. Huang, and W. Gao, "Ram: A region-aware deep model for vehicle re-identification," in *Proceedings of the International Conference on Multimedia and Expo*, 2018, pp. 1–6.

[52] L. V. Der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[53] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, 2020.

[54] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," 2017, pp. 1–11.

[55] J. Chen, T. Ma, and C. Xiao, "Fastgcn: fast learning with graph convolutional networks via importance sampling," *arXiv preprint arXiv:1801.10247*, 2018.

Hongchao Li received the B.Eng. degree in software engineering and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 2017 and 2022, respectively. He is currently a Lecturer with the School of Computer and Information, Anhui Normal University. His current research interests include person/vehicle re-identification and multi-modal learning.

AiHua Zheng received her B.Eng. degrees and finished her Master-Doctor combined program in computer science and technology from AnHui University of China in 2006 and 2008 respectively. And received her Ph.D degree in computer science from the University of Greenwich of UK in 2012. She is currently an associated professor in Artificial Intelligence at 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.

Liping Sun received the Ph.D. degree from Anhui Normal University, Wuhu, China, in 2015. She is currently a Professor with the School of Computer and Information, Anhui Normal University. Her current research interests include data mining and intelligent computing.

Yonglong Luo received his PhD degree from the School of Computer Science and Technology, University of Science and Technology of China in 2005. Since 2007, he has been a professor in School of Computer and Information, Anhui Normal University. Currently, he is the PhD supervisor of Anhui Normal University. He is the Director of Anhui Provincial Key Laboratory of Network and Information Security. His research interests include information security and spatial data processing.