Article:
Xu, Zheming, Wei, Lili, Lang, Congyan et al. (2022) SSR-Net: A Spatial Structural Relation Network for Vehicle Re-identification. ACM Transactions on Multimedia Computing Communications and Applications. 3578578. ISSN: 1551-6865

https://doi.org/10.1145/3578578

SSR-Net: A Spatial Structural Relation Network for Vehicle Re-identification

ZHEMING XU, LILI WEI, CONGYAN LANG*, SONGHE FENG, and TAO WANG, the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, China
ADRIAN G. BORS, University of York, The United Kingdom
HONGZHE LIU, Beijing Key Laboratory of Information Service Engineering, Beijing Union University, China

Vehicle re-identification (Re-ID) represents the task aiming to identify the same vehicle from images captured by different cameras. Recent years have seen various feature learning based approaches merely focusing on feature representations including global features or local features to obtain more subtle details to identify highly similar vehicles. However, few such methods consider the spatial geometrical structure relationship among local regions or between the global and local regions. By contrast, in this study, we propose a Spatial Structural Relation Network (SSR-Net) which explores the above-mentioned two kinds of relations simultaneously to learn more discriminative features by modeling the spatial structure information and global context information. In this paper we propose to adopt a Graph Convolution Network (GCN), for modeling spatial structural relationships among characteristic features. The GCN model aggregating the local and global features is shown to be more discriminative and robust to several car image transformations. To improve the performance of our proposed network, we jointly combine the classification loss with metric learning loss. Extensive experiments conducted on the public VehicleID and VeRi-776 datasets validate the effectiveness of our approach in comparison with recent works.

CCS Concepts: • **Computing methodologies** → **Object identification**; *Image representations.*

Additional Key Words and Phrases: Vehicle re-identification, Graph convolution network, Attention Mechanism, Deep learning

1 INTRODUCTION

Vehicle re-identification (vehicle Re-ID), aiming to identify the same vehicle from a gallery of images captured by disjoint cameras, has attracted much attention in the computer vision community in recent years. Vehicle Re-ID has been widely applied in urban surveillance scenarios, *e.g.*, traffic management, video surveillance, and intelligent security. It plays an important role in retrieving suspicious vehicles from giant surveillance data to save labor costs and improve efficiency. However, Vehicle Re-ID still remains a challenging task due to subtle inter-class variation (*i.e.*, variation between different vehicles) and enormous intra-class variation (*i.e.*, variation among the same vehicle in different images) caused by severe changes in the image acquisition variation, such

*Corresponding Author.

Authors' addresses: Zheming Xu, 21112016@bjtu.edu.cn; Lili Wei, 20112014@bjtu.edu.cn; Congyan Lang, cylang@bjtu.edu.cn; Songhe Feng, shfeng@bjtu.edu.cn; Tao Wang, twang@bjtu.edu.cn, the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, China, 100044; Adrian G. Bors, University of York, York, The United Kingdom, adrian.bors@york.ac.uk; Hongzhe Liu, Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing, China, liuhongzhe@buu.edu.cn.

1551-6857/2022/12-ART $15.00
https://doi.org/10.1145/3578578

as heavy occlusion of vehicles, illumination conditions, and the variation in the visual appearance caused by changes in the viewpoints of the cameras acquiring the vehicle images.

To address the above challenges, there are two different categories of approaches, *i.e.*, distance metric learning based approaches and feature learning based approaches. The first group emphasizes how to map all samples into a suitable latent subspace and relies on different loss functions to constrain the inter- and intra-class distances. By contrast, the other group focuses on how to learn more discriminative features to enhance the performance and robustness, as shown in Fig. 1. Early works of this group [1, 35, 77] merely extract global features by adopting different deep convolutional neural networks or different loss functions, as shown in Fig. 1 (a). However, these approaches lack exploiting subtle cues in local regions (*e.g.*, vehicle logos, vehicle lights, annual inspection insurance stickers, and interior decorations) , which are more discriminative in identifying highly similar vehicles with similar colors and vehicle types. On account of this, some approaches [16, 40, 63] turn to extract global and local feature simultaneously by adopting several separate methods embedding the local features as complementary information for global features to distinguish similar vehicles between different classes, as shown in Fig. 1 (b). However, these approaches ignore intrinsic relations within an image, which can be helpful for learning more spatial structure information. To this end, recent advances [39, 90] make progress in exploring the intrinsic relations among local regions, as shown in Fig. 1 (c). However, prior studies do not consider the relationships between each local region and the global area, which could be helpful to improve the global context awareness of each local region, further making the extracted features more discriminative and robust.

In this work, inspired by the capabilities of the Graph Convolution Networks (GCN) [2, 12, 30], we propose a **S**patial **S**tructural **R**elation **Net**work (SSR-Net) to address the challenges posed by vehicle Re-ID. Compared with previous research studies, we attempt to model two kinds of relations simultaneously, *i.e.*, the relations among local regions to learn spatial structural information, and the relations between each local region and the global relevant information to enhance the context awareness of local regions, as shown in the blue lines and red lines in Fig. 1 (d). By learning these relations, the SSR-Net can learn more discriminative features with spatial structure information and global context information. Concretely, the SSR-Net is composed of three branches, *i.e.*, **G**lobal **B**ranch (GB), **A**ttention **B**ranch (AB) and **R**elation **B**ranch (RB), respectively. Among them, GB is used to extract the global feature representation of each vehicle image. Based on the global feature, AB further relies on two attention mechanisms to generate more discriminative feature representations, which usually pay more attention to subtle differences in local regions. Subsequently, RB, the core branch of SSR-Net, aims to model the structure relationships among local regions or between the global region and local regions. Specifically, we first construct a spatial geometrical structure graph by taking the global feature as a global node and the five local feature patches cropped from AB as local nodes. Each local node is directly associated with its adjacent local nodes. To avoid over-smoothing issues when directly connecting global and local nodes, we introduce a learnable token node to indirectly associate the global node and local nodes. Based on such a spatial structural graph, RB employs a GCN module to learn more discriminative structure features by incorporating global and local features. After SSR-Net learning three different kinds of feature representations for each vehicle, in order to further boost the performance, we utilize a combination of classification loss and metric learning loss for the training of the whole SSR-Net.

We conduct extensive experiments on two Re-ID benchmark datasets, and experimental results demonstrate the effectiveness of the proposed SSR-Net. The main contributions of the proposed approach are summarized as follows:

- We propose a Spatial Structural Relation Network (SSR-Net) for vehicle Re-ID by processing three branches and jointly exploiting the relationships among the global and local vehicle image representations.

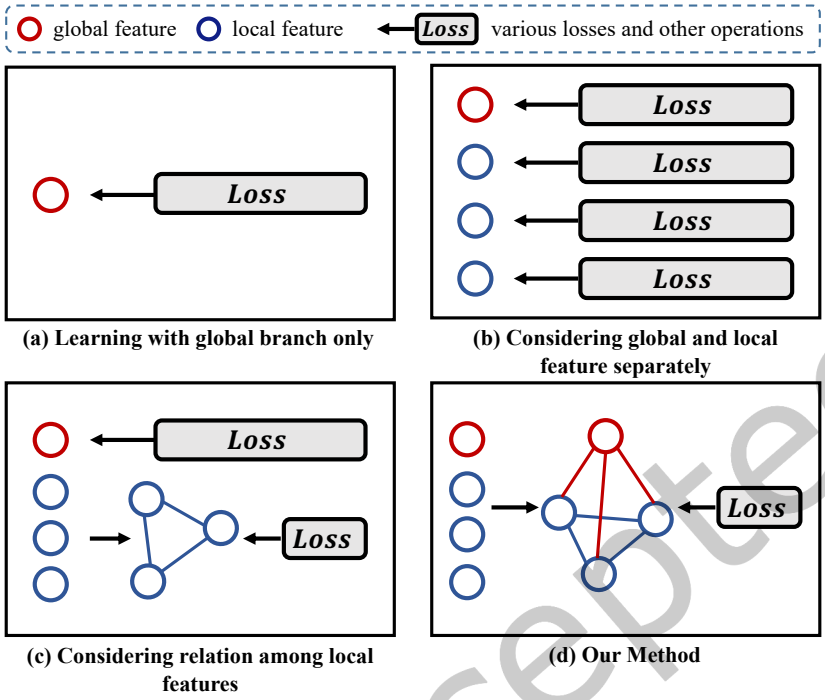


Fig. 1. The illustration of comparison between previous re-id methods and ours. (a) approaches merely extracting global features; (b) approaches separately extracting global and local features; (c) approaches with relations among local regions; (d) the proposed SSR-Net exploring spatial relations among local regions and the relations between global areas and local regions. The blue lines indicate the relations among local regions, and the red lines represents relations between each local region and the global representation.

- We model the relationships among local features and between the global feature and local features in the form of a spatial geometrical structure graph, and utilize a GCN module to conduct message-passing on the graph topology to extract spatial structural relation representations.
- Extensive comparison experiments and ablation studies on two benchmark datasets demonstrate the effectiveness of the proposed SSR-Net.

Our preceding work is described in [69]. In this work, our crucial improvements include: (1) We proposed an attention branch to obtain detailed features which are more discriminative. (2) We reconstruct the graph by introducing a learnable token node to learn spatial structure information and global context information simultaneously, while avoiding over-smoothing. (3) In each branch, we jointly learn the triplet loss and cross-entropy loss.

2 RELATED WORKS

In this section, we first briefly revisit the most relevant works in the field of Re-ID in Sec. 2.1, followed by a review of the work in Graph Convolutional Network in Sec. 2.2. Finally, in Sec. 2.3, we briefly introduce the research on attention mechanisms.

2.1 Vehicle Re-ID

In recent years, various research studies have addressed the field of vehicle Re-ID, and these works can be separated into two categories: metric learning based approaches and feature learning based approaches.

Metric learning based approaches. The metric learning methods [1, 5, 35, 37, 41, 77] focus on learning a latent embedding in order to decrease the intra-class variance and increase inter-class variance to improve the discrimination. Depending on the number of network inputs, metric learning can be classified as contrastive loss [15, 37], triplet loss [51], and quadruplet loss [21]. In the vehicle Re-ID community, the triplet loss and its variants [77] are now the most prevalent metric learning loss functions employed.

Feature learning based approaches. Over the last few decades, most of the existing research has aimed to extract more discriminative features. These efforts may be divided into two categories, namely hand-crafted features extraction and deep learning based features extraction. In an early age, researchers designed and utilized different feature extractors [33, 46, 58, 76] to obtain hand-crafted features. For instance, Liu *et al.* [36] utilized SIFT descriptor [46] and CN (Color Name) descriptor [58] to extract the texture and color feature of local regions. Liao *et al.* [33] presented a feature descriptor termed Local Maximal Occurrence (LOMO) to cope with the person Re-ID problem. However, the method based on hand-crafted features has its drawbacks. First, low-level semantic features, such as color and texture, are vulnerable to the environment (*e.g.*, lighting, shooting angles and resolution). Furthermore, these methods demand a lot of time and effort from researchers to design effective feature descriptors, which is inefficient.

After the great success and widespread of AlexNet and the following development of Convolutional Neural Networks (CNN), deep learning methods have shown their great power in various applications. The task of vehicle re-identification is no exception [35, 44, 65, 79, 80]. Early research methods mainly focus on extracting the global feature of the vehicle images. For instance, Guo *et al.* [13] propose a structured feature embedding method that designed coarse-grained to fine-grained ranking loss to extract more discriminative global features. However, these methods neglect to consider local regions which contain rich detailed information that helps deal with challenges like subtle inter-class variance. In order to deal with this challenge, He *et al.* [16] propose a network with two modules, namely the Global Module (GM) and the Local Module (LM). In LM, three kinds of local features (*i.e.*, rear windows, headlights, license plate logos) are obtained by the YOLO detection algorithm [49]. Besides dividing overlapped parts to obtain local features, Liu *et al.* [40] also introduce two extra relevant attributes (*i.e.*, colors, and vehicle models) to further improve the performance of the method. He *et al.* [18] propose an approach that introduces a two-branch network to extract the appearance of vehicles, as well as a license plate Re-ID network to capture the contexts of license plate images to further boost the performance of vehicle Re-ID.

However, most of these methods only pay attention to obtaining visualized feature expression while ignoring considering the structural relationship of vehicle representation. Nevertheless, recently some research studies have considered the structural relationship of local regions of vehicles. For instance, Liu *et al.* [39] build a Multi-grained Vehicle Parsing (MVP) dataset by constructing a semantic graph to further extract the structural relationship among local regions. Zhu *et al.* [90] propose a Structured Graph Attention network (SGAT) to exploit the structural relationship among local landmarks of vehicles. However, these methods only exploit the relationship among local features while not considering the global representation or using this separately. In this paper, we consider the two relationships simultaneously.

2.2 Graph Convolutional Network (GCN)

GCN. In real-life, plenty of structured data exists in a graph manner, such as social networks, protein structures, transportation networks and the World Wide Web. Although the convolutional network has gained a profound

reputation in processing images, audios and sentences, it is not appropriate to be used for certain data representations. Graph Convolution Network (GCN) is one of the methods that can process graph representations. The study of GCN mainly focuses on the propagation and aggregation of information from neighboring nodes. In terms of feature space, GCN can be generally divided into two categories: spectral-domain and spatial-domain. The spectral GCN [2, 7, 19, 32, 75] define convolution operations on the graph by using Fourier or Laplace transform, while the spatial GCN [12, 50, 74] aggregate each central node and its neighboring nodes in the graph directly by defining various aggregation functions. GCN has gained much attention in recent years and has been applied in various tasks such as skeleton-based action recognition [70, 71], zero-shot learning [61], few shot learning [11], social relation recognition [38], point clouds processing [62], *etc.*

GCN in Re-ID Tasks. GCN has also been applied in the Re-ID community. For the person Re-ID task, Jiang *et al.* [25] propose a Part-based Hierarchical Graph Convolutional Network (PH-GCN) which builds a hierarchical graph to represent the pairwise relationships across distinct regions and utilizes a GCN to extract the structural feature of images. In terms of the video-based person Re-ID task, Yang *et al.* [72] propose a Spatial-temporal Graph Convolutional Network (STGCN) to learn the temporal relationship between distinct frames and the spatial relationship inside a given frame. Ji *et al.* [24] develop a Meta Pairwise Relationship Distillation (MPRD) method, in which a GCN module predicted the pseudo labels of sample pairs to deal with the unsupervised person Re-ID task. As for the vehicle Re-ID task, Liu *et al.* [39] propose a Parsing-guided Cross-part Reasoning Network (PCRNet) to extract part-level feature representation and then adopt a GCN module to explore the relations between various parsing-guided parts.

2.3 Attention Mechanism

Attention mechanisms [10, 22, 48, 53, 66] have gained great popularity in recent years.They have been utilized in a variety of domains, including action recognition [52], image caption generation [68], human pose estimation[6], and image classification [22, 60, 64, 67, 78]. Specifically, Hu *et al.* [22] propose squeeze-and-excitation (SE) module to adjust the weights of channels. Based on the SENet, the CBAM [64] considers spatial attention module and channel attention module simultaneously. Vaswani *et al.* [59] propose the transformer, which is designed entirely based on the self-attention mechanism and inspires the development of ViT [9], Swin Transformer [43] and other variants.

3 METHODOLOGY

In this section, we provide the overview of the proposed re-ID framework in Sec. 3.1. Then the three branches, *i.e.*, Global Branch, Attention Branch, and Relation Branch are elaborated in Sec. 3.2, Sec. 3.3, and Sec. 3.4, respectively. Finally, in Sec. 3.5, we introduce the joint learning of classification loss with metric learning loss.

3.1 Framework Overview

Given input images $X = \{x_1, x_2, \ldots, x_B\}$ and the ground truth labels $Y = \{y_1, y_2, .., y_M\}$, where B and M represent the number of images in a mini batch and the total number of classes respectively, the aim of the **S**patial **S**tructural **R**elation **Net**work (SSR-Net) is to extract robust and discriminative features for vehicle Re-ID. The overall pipeline of our proposed SSR-Net is briefly illustrated in Fig. 2, where SSR-Net consists of three components, namely Global Branch (GB), Attention Branch (AB), and Relation Branch (RB). Specifically, GB adopts a simple yet effective backbone network $\mathcal{F}(\cdot)$ to generate the global representation (global feature map $F_G \in \mathbb{R}^{B\times C\times H\times W}$ and global feature vector f_G) for input images X, where B, C, H and W are the batch size, channel, height and width of F_G, respectively. Later, AB takes F_G as input, learns important subtle features via the Attention Module (AM), and derives the attention feature map which is denoted as F_{Attn} in Fig. 2 (b). Subsequently, as depicted in Fig. 2 (c), taken F_{Attn} as input, RB first crops F_{Attn} into a local feature map set $F_L = \{F_{ul}, F_{ur}, F_{mid}, F_{dl}, F_{dr}\}$ with five local

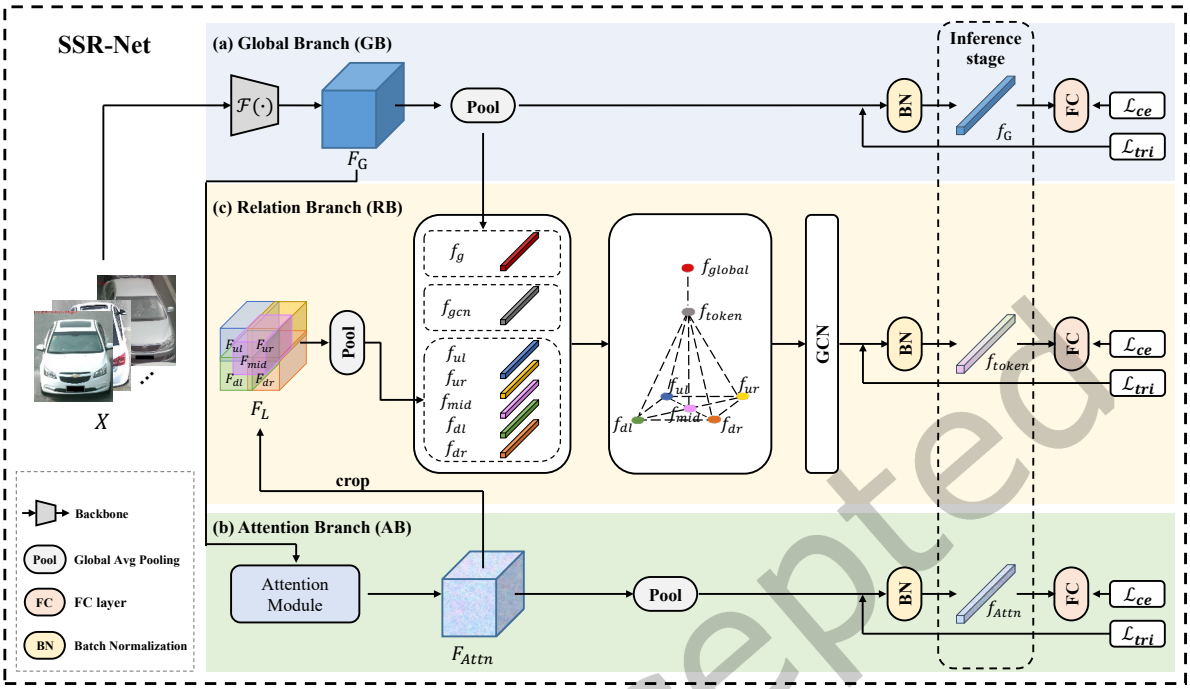


Fig. 2. The framework overview illustration of the proposed SSR-Net. (a) **g**lobal **b**ranch (GB) embeds the input images into latent space. (b) **a**ttention **b**ranch (AB) takes global feature map as input and conduct attention operations to obtain attention maps. (c) **r**elation **b**ranch (RB) organizes global and local representations in the form of graph, and utilizes a GCN module to incorporate the geometrical structural information and derive discriminative features. Best viewed in color.

feature maps, and then derives five different local representations (denoted as $f_l = \{f_{ul}, f_{ur}, f_{mid}, f_{dl}, f_{dr}\}$) which have low spatial correlation in the original image. Later, considering f_l, f_g and a randomly initialized feature vector f_{token} as node feature embeddings, RB models the spatial structural relation among local regions and that between global and local areas using a graph representation. Afterward, a GCN module is utilized to make message-passing on the graph topology to derive an updated representation f_{token} which integrates both the global information and spatial local information. Note that during the inference stage, concatenation of extracted f_g, f_{token} and f_{Attn} is denoted as the final representation for vehicle Re-ID.

3.2 Global Branch

As depicted in Fig. 2 (a), GB is designed to derive global representation of the input images X. In the branch, a CNN $\mathcal{F}(\cdot)$ (such as ResNet-50 [17]) is trained to infer the latent space corresponding to a given image X. After obtaining the feature map F_G, there are two data flows. In one flow, global feature vector F_G is derived after operations with global average pooling (GAP) and batch normalization (BN). In the other flow, F_G is taken as the input of AB as explained in Sec. 3.3 to help with subsequent attention operations.

3.3 Attention Branch

As aforementioned, subtle detail information in local regions could effectively help re-identify the vehicle. To this end, we adopt a light-weight attention module CBAM, proposed in [64], to obtain local attention maps. Denote

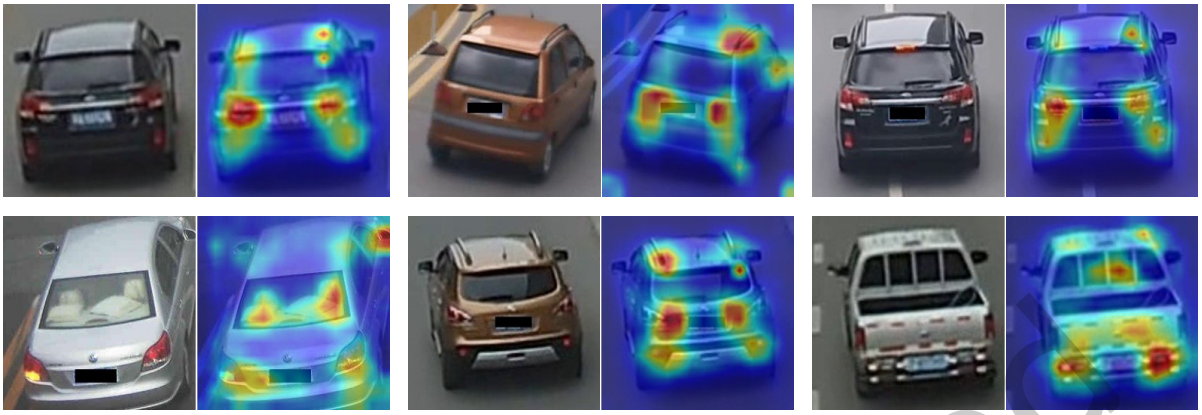


Fig. 3. Visualization of attention maps. Warm color represents the high significance of its covered region.

the channel attention map obtained from CBAM as $M_c \in \mathbb{R}^{C\times1\times1}$ and the spatial attention map as $M_s \in \mathbb{R}^{1\times H\times W}$. Given the global feature map F_G as input, the attention map F_{Attn} is computed as follows:

$$\begin{aligned} F_c &= M_c(F_G) \otimes F_G, \\ F_{Attn} &= M_s(F_c) \otimes F_c, \end{aligned} \tag{1}$$

where $\otimes$ represents element-wise multiplication. F_{Attn} provides detailed cues for vehicle Re-ID, and is taken as the input of the Relation Branch in Fig. 2 (c).

3.4 Relation Branch

Feature Extraction. In general, the middle part of the image contains rich local details because the vehicle is cropped using object detection algorithms. Meanwhile, the vehicle is a rigid object and the information of a vehicle is defined by its parallelepipedic geometrical structure. As illustrated in Fig. 3, the majority of the focus areas are located in the four corners of the segmented vehicle object. Therefore, we denote the four corners and the middle region of the vehicle image as the local feature map set $F_L = \{F_{ul}, F_{ur}, F_{mid}, F_{dl}, F_{dr}\}$, in which $F_i \in \mathbb{R}^{B\times C\times H_L\times W_L}$. H_L and W_L represent the height and length of each local feature map, and are computed as follows:

$$\begin{aligned} H_L &= \lfloor \lambda \times H \rfloor, \\ W_L &= \lfloor \lambda \times W \rfloor, \end{aligned} \tag{2}$$

where $\lfloor\rfloor$ represents floor operator and $\lambda \in (0, 1)$ is a region ratio parameter. For improving the robustness of the vehicle representation we consider overlaps between the selected regions by considering $\lambda > \frac{1}{2}$. Considering alignment error and pose variation, we set λ as $\frac{2}{3}$ to define overlapping regions in the vehicle representation data.

Subsequently, in each branch, feature maps are embedded as feature vector representations, denoted as $f_G, f_{ul}, f_{ur}, f_{mid}, f_{dl}, f_{dr}$. To better integrate both the global and spatial local information, as well as avoid over-smoothing, we further introduce a token feature representation f_{token} that is initialized randomly.

Graph Construction. In recent years, some existing methods concatenate local features directly, ignoring the latent relationships between them. Some approaches consider such intrinsic relation among local parts whereas ignoring the spatial structural relationship between local and global features at the same time. To this end, we formulate these two relations in the form of a graph depicted in the middle of Fig. 2.

Let $G(V, E)$ represents the constructed graph. $G(V, E)$ is composed of seven nodes including five local nodes, one global node and one token node which is initialized randomly. Therefore, the node set V is defined as $V = \{v_G, v_{ul}, v_{ur}, v_{mid}, v_{dl}, v_{dr}, v_{token}\}$. As illustrated in Fig. 2, edge $e_{ij} = 1$ when v_i and v_j are connected. The adjacency matrix $A \in \mathbb{R}^{n \times n}$ can be then formulated as follows,

$$a_{ij} = \begin{cases} 1 & e_{ij} = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $a_{ij} \in A$ and n is the number of nodes. At last, the adjacency matrix of $G(V, E)$ is denoted as $\widetilde{A} = A + I$, where I is the identity matrix.

Feature Update. Following the work of Kipf [30], we employ a l th-layer GCN module (l is set as 2 in this paper) to extract structural features. To be more specific, the output of the $l + 1$ th layer of the GCN module can be represented as,

$$H^{(l+1)} = \sigma(\widetilde{D}^{-1}\widetilde{A}H^{(l)}W^{(l)}), \tag{4}$$

where $\widetilde{D} \in \mathbb{R}^{n \times n}$ is the degree matrix of $\widetilde{A}$. The activation function $\sigma(\cdot)$ is $ReLU(\cdot)$. The initial input of the GCN module $H^{(0)}$ is consist of the feature vector of the seven nodes. In the end, the updated token feature f_{token} represents the output of RB, containing spatial structural relationships.

3.5 Loss Functions

Classification Loss. The vehicle Re-ID task can be treated as a multi-class classification problem. Therefore, given sample images X, we employ the cross-entropy loss,

$$\mathcal{L}_{ce} = -\frac{1}{B}\sum_{b=1}^{B}\sum_{m=1}^{M} y_{im} log(p_{im}), \tag{5}$$

where B is the number of samples in a mini batch, and M is the number of classes. The ground truth y_{im} equals 1 when the i th sample belongs to the m th class of vehicles, and p_{im} represents the predicted probability that an image belongs to the m th vehicle class. In training stage, feature vectors obtained after the fully connected layers in the three branches are supervised by the cross-entropy loss which can be denoted as $\mathcal{L}_{ce}^{G}$,$\mathcal{L}_{ce}^{SSR}$, $\mathcal{L}_{ce}^{Attn}$ respectively.

Metric Embedding Loss. To learn more discriminative representations, we employ the triplet with hard example mining [20] which selects challenging samples in each batch as positive and negative samples. The loss is computed as follows:

$$\mathcal{L}_{triHard} = \sum_{p=1}^{P}\sum_{k=1}^{K}((\max_{i=1,\dots k} D(x_p^k, x_p^i) - \min_{\substack{n=1,\dots P \\ j=1,\dots k \\ n \neq p}} D(x_p^k, x_n^j) + \alpha))_+, \tag{6}$$

where x_p^k denotes the k th image of the p th vehicle, and $D(\cdot)$ represents the distance metric function. Function $(\cdot)_+$ denotes $max(0, \cdot)$, and hyper-parameter α represents a margin between positive and negative pairs. Feature vectors obtained before the batch normalization layers in the three branches are supervised by the enhanced triplet loss whose components are denoted $\mathcal{L}_{triHard}^{G}$, $\mathcal{L}_{triHard}^{SSR}$ and $\mathcal{L}_{triHard}^{Attn}$.

Total Loss. The triplet loss focuses on the similarity distance learning, whereas the cross-entropy loss seeks to find a hyper-plane to classify data. In this work, we learn the triplet loss and the cross-entropy loss simultaneously. A batch normalization (BN) layer is used before the fully connected layer. For our proposed methods, the total loss is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce}^{G} + \mathcal{L}_{ce}^{SSR} + \mathcal{L}_{ce}^{Attn} + \beta(\mathcal{L}_{triHard}^{G} + \mathcal{L}_{triHard}^{SSR} + \mathcal{L}_{triHard}^{Attn}), \tag{7}$$

where β represents the loss balance parameter.

4 EXPERIMENTS

In this section, we evaluate the proposed SSR-Net through extensive experiments. To be more specific, we first introduce two public datasets and the evaluation metrics in Sec. 4.1. Later, implementation details are elaborated in Sec. 4.2. Then we conduct extensive ablation studies to demonstrate the effectiveness of essential components of SSR-Net in Sec. 4.3. In Sec. 4.4, we conduct various experiments to explore the impact of different hyper-parameters. Subsequently, we validate the proposed method on two benchmarks and compare the obtained results with the state-of-the-art methods. Finally, we visualize the retrieval results in two retrieval datasets and show the superiority of the proposed SSR-Net in Sec. 4.7.

4.1 Datasets and Evaluation Metrics

Datasets. *VehicleID* [35] dataset is one of the common datasets in vehicle Re-ID tasks. It is also known as the PKU VehicleID dataset, which is collected and published by the National Engineering Laboratory for Video Technology (NELVT) of Peking University in 2016. All images in the dataset are derived from real surveillance data captured from traffic cameras. VehicleID contains a total of 221,763 images from 26,267 vehicles. On average, the number of images per vehicle is 8.44, and each vehicle has at least two images. Therefore, the dataset is suitable for the vehicle re-identification task. In order to protect the privacy of vehicle owners, the dataset obscures the license plates information of all images with black masks. Based on the size of the dataset, VehicleID includes three test subsets: VehicleID-800, VehicleID-1600, and VehicleID-2400. There are 6493 images from 800 vehicles in VehicleID-800, 13777 images from 1600 vehicles in VehicleID-1600, and 19777 images from 2400 vehicles in VehicleID-2400.

VeRi-776 [37] is collected from traffic data captured by 20 different cameras. The dataset contains 49360 images with 776 vehicles. In addition, VeRi-776 is annotated with information such as bounding box, vehicle type, and color, and supplemented with spatial-temporal information such as the distance between various cameras. In VeRi-776, there are 37781 images of 576 identities used for training, while the rest 11579 images of 200 vehicles constitute the test set.

Evaluation Metrics. To validate the effectiveness of our proposed method, we adopt two evaluation metrics as the previous work does: Cumulative Matching Characteristics (CMC) and mean average precision (mAP). CMC curve reflects retrieval accuracy. Generally, CMC@K is often adopted, where K represents the hit accuracy of the top K positions. The formulation of CMC@K is defined as follows:

$$CMC@K = \frac{\sum_{q=1}^{Q} gt(q, K)}{Q}, \tag{8}$$

where q refers to the q th probe image, and Q refers to the total number of probes. Ground truth $gt(q, K) = 1$ when the q th probe is in the top-K of the rank list, and otherwise, $gt(q, K) = 0$.

The CMC curve is suitable for single-gallery-shot, and CMC@1, also known as Rank-1, is an extremely significant metric especially in application scenarios. The evaluation metric mean average precision (mAP) is another appropriate metric especially when there are multiple query results in the gallery. The equation of mAP is computed as follows:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{9}$$

where AP refers to the average precision, and Q refers to the total number of probe images.

Table 1. Evaluation of the effectiveness of the crucial components within SSR-Net on datasets. "RB w/o GN" denotes that the constructed graph ignores the global node and only formulates the intrinsic relation among local nodes. "RB w/o token" denotes constructing a graph with global and local nodes connected directly. The best results are indicated with bold.

Methods	VeRi-776		VehicleID-800		VehicleID-1600		VehicleID-2400	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
GB	93.31	71.54	77.74	83.97	74.37	79.90	71.67	77.00
+AB	95.23	76.23	81.73	87.41	75.96	82.65	74.75	81.11
+RB (w/o GN)	94.87	74.97	82.73	88.62	77.04	83.49	75.94	81.79
+RB (w/o token)	95.05	75.09	82.72	88.22	76.96	83.36	74.38	80.93
+RB	95.17	76.07	83.01	88.56	78.85	84.67	77.08	82.84
+AB+RB (w/o GN)	95.95	77.72	82.38	88.03	79.07	84.98	75.99	81.89
+AB+RB (Our Method)	**96.36**	**78.32**	**83.08**	**89.07**	**79.51**	**85.13**	**77.44**	**83.12**

4.2 Implementation Details

Network setting. In this work, ResNet-50 is used for extracting basic features of input images, and is pre-trained on ImageNet [8]. The input images are resized to 256×256 and we apply random flipping and cropping for data augmentation. The dimension of each local feature is set as $\mathbb{R}^{B\times 2048}$ where B represents the batch size. In all experiments, B is set as 32, in which the number of the IDs per batch is 8 while the number of images per ID is 4. Note that the token node is randomly initialized with a dimension of $\mathbb{R}^{1\times 2048}$. The dimension of the output feature of RB is set as $\mathbb{R}^{B\times 512}$.

Hyper Parameters. All experiments are implemented using PyTorch, and are conducted with 4 12G NVIDIA TITAN XP GPUs. In each mini-batch, Stochastic Gradient Descent (SGD) optimizer is adopted to train the network with the weight decay of 5×10^{-4}. The initial learning rate is set as 1×10^{-4} and starts decaying at epoch 15 by cosine annealing. The warm-up strategy is employed to the learning rate in the first 5 epochs. The margin α of metric embedding loss in Eq. (6) is set as 0.5. The value of hyper-parameters λ, β and l are studied in Sec. 4.3 and are finally set as $\frac{2}{3}$, 1 and 2, respectively.

4.3 Ablation Studies

Overall, our proposed model is made up of three parts, namely the global branch (GB), the attention branch (AB), and the relation branch (RB). GB extracts global features through the common convolutional network, which is also considered as the baseline. By introducing attention mechanisms, AB identifies crucial local features. The RB defines the structure information as a graph with one global node, one token node, and five local nodes. Then we employ the graph convolutional network to explore the structural relationship among local parts and that between local and global features. In order to fully validate the effectiveness of each branch in the proposed method, we conduct a series of ablation experiments on both VehicleID and VeRi-776 datasets.

The Effectiveness of each component. As illustrated in Table 1, on the basis of "GB", by incorporating AB, "+AB" achieves better performance over "GB", with Rank-1 accuracy and mAP increasing by 1.92% and 4.5% on VeRi-776, respectively. On VehicleID-2400, "+AB" gains superior performance, with Rank-1 accuracy raised by 3.08% while mAP increased by 4.11%. These results demonstrate that the attention module further extracts useful local features and would enhance the performance of GB. In terms of the RB, we observe that "+RB" gains at least 4.43% higher mAP performance over the baseline. Such results demonstrate the effectiveness and robustness of the proposed method. According to Table 1, "+AB+RB" beats the baseline by 5.77% and 6.12% in Rank-1 accuracy and mAP on VehicleID-2400, respectively. A similar trend also appears on the results achieved on the VeRi-776

Table 2. Evaluation on different attention-based methods on VeRi-776 dataset where * denotes that the same base networks are used.

Methods	VeRi-776		
	Rank-1	Rank-5	mAP
Bilinear CNN* [34]	87.57	95.05	59.39
Spatial Transformer Network* [23]	95.10	97.97	72.70
SE-Block* [22]	95.49	98.09	75.96
CBAM [64]	**96.36**	**98.39**	**78.32**

Table 3. Evaluation on different part-based methods on VehicleID and VeRi-776. * denotes that the same base networks are used.

Method	VeRi-776		VehicleID-800		VehicleID-1600		VehicleID-2400	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
PCB*[56]	95.59	78.04	77.99	85.21	73.43	80.94	71.43	78.97
SSR-Net	**96.36**	**78.32**	**83.08**	**89.07**	**79.51**	**85.13**	**77.44**	**83.12**

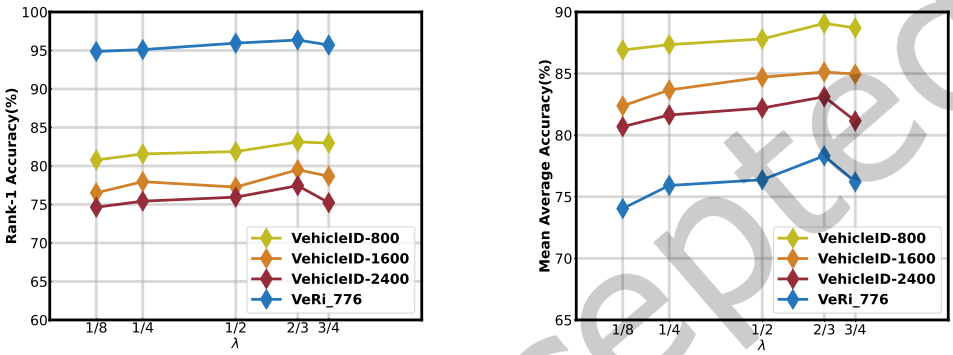
dataset. Furthermore, we observe that the combination of all modules achieves better performance than when each individual module is considered, demonstrating the effectiveness of our proposed approach.

The Effectiveness of different ways of graph construction. In the following, we test various model choices, for the relationships between local and global features in the GCN model for re-ID. The configurations considered "+RB (w/o GN)" and "+AB+RB (w/o GN)", where the global node is abbreviated as "GN". Besides, to validate the effectiveness of the token node, we also create a model variant called "+RB (w/o token)" that builds a graph without a token node, *i.e.,* connecting global and local nodes directly. From Table 1, it can be observed that approaches considering the global-local relation exceeds both their variants. To be more explicit, on the VeRi-776 dataset, "RB" improves "+RB (w/o GN)" by 0.3% and 1.3% on Rank-1 and mAP respectively. "RB" also results in 0.98% higher performance on mAP compared with "+RB (w/o token)", showing the effectiveness of the token node. In the case of VehicleID-2400, "RB" holds a larger Rank-1 accuracy and mAP of 1.14% and 1.05%, respectively. For the comparison of "+AB+RB" and "+AB+RB (w/o GN)", we similarly observe that the former approach outperforms its variant by 0.6%, 1.04%, 0.15% and 1.23% mAP on VeRi-776, VehicleID-800, VehicleID-1600, VehicleID-2400, respectively.

The Effectiveness of different attention-based and part-feature based methods. In Table 2 and Table 3, we further compare the performance of different attention-based methods and part-feature-based methods, respectively. As illustrated in Table 2, we consider Bilinear CNN[34], SN-Block[22] and CBAM [64] on VeRi-776. Besides, following PAN[83] and TAMR[14], we also validate the effectiveness of the Spatial Transformer Network (STN)[23] module. Note that we directly utilize STN after input images for simplicity. Results in Table 2 show that by considering spatial and channel attention simultaneously, CBAM achieves the best results. Therefore, CBAM is used as our attention-mechanism method in the following experiments. Table 3 shows the performance of different part-feature-based methods on VeRi-776. Compared to the conventional PCB[56], our method which crops features through patches, achieves better results.

Table 4. Rank-1 accuracy and mAP results obtained with different values of l on VehicleID and VeRi-776.

l	VeRi-776		VehicleID-800		VehicleID-1600		VehicleID-2400	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1	96.07	77.12	82.5	88.04	79.08	84.65	76.08	81.81
2	**96.36**	**78.32**	**83.08**	**89.07**	**79.51**	**85.13**	**77.44**	**83.12**
3	95.47	76.97	81.93	87.77	78.02	84.20	75.21	81.26



(a) Rank-1 accuracy results. (b) mAP results.

Fig. 4. Rank-1 accuracy and mAP results obtained with different values of λ.

4.4 Impact of Hyper-parameters

In this subsection, we evaluate the impact of three hyper-parameters in SSR-Net, which are the depth of GCN l, the ratio λ in Eq. (2) controlling the size of local regions and the loss balance parameter β.

Impact of the depth l for GCN. As shown in Fig. 2, the maximum depth of the constructed graph structure is 2. Based on this, we consider three different values of l to validate the model performance. The experimental results are shown in Table 4. The best performance is obtained when $l = 2$. Meanwhile, a three-layer GCN has the poorest performance, with 1.19% lower mAP on VeRi-776 compared with model setting l to 2. On the basis of this phenomenon, we analyze that over-fitting occurs when the depth is over the maximum depth of the graph, which in turn decreases the performance of the model.

Impact of the crop ratio λ. In Sec. 3.4 we consider a hyper-parameter λ controlling the size of local areas. Fig. 4 depicts the experimental results of this hyper-parameter on VehicleID and VeRi-776s. Overall, the Rank-1 accuracy is not sensible to λ as the variation trend in Fig. 4 (a) is not obvious. However, for another evaluation metric, Fig. 4 (b) indicates an obvious variation trend. To be more specific, on both datasets, the model has poor performance on Rank-1 accuracy and mAP when λ is set as $\frac{1}{8}$, then its performance increased sharply as λ varies from $\frac{1}{8}$ to $\frac{2}{3}$, and reached the best performance when $\lambda = \frac{2}{3}$. For higher values, the model's performance decreases dramatically. The results demonstrate that a small λ setting leads to a rather weak expression for re-identification, while a larger λ results in local feature maps similar to the global feature maps, causing over-smoothing issues during GCN processing. In the following comparison and ablation studies, we consider λ set as $\frac{2}{3}$.

The impact of loss balance parameter β. To validate the impact of balance parameter β in Eq. (7), we test the value of β from 0 to 1.5 with a step of 0.2 or 0.3, and conduct distinctive experiments, as reported in Table 5. As

Table 5. Rank-1 accuracy and mAP results obtained with different values of β on VehicleID and VeRi-776.

β	VeRi-776		VehicleID-800		VehicleID-1600		VehicleID-2400	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
0	94.04	73.43	78.90	85.36	75.32	81.44	74.44	80.14
0.2	94.40	74.32	79.40	85.99	78.78	84.58	75.30	81.56
0.5	94.34	74.27	81.40	87.49	78.47	84.43	75.39	81.51
0.8	94.70	75.02	81.85	87.96	78.47	84.61	77.04	82.88
1.0	**96.36**	**78.32**	**83.08**	**89.07**	**79.51**	**85.13**	**77.44**	**83.12**
1.2	94.22	73.58	81.49	87.70	78.17	84.17	76.19	82.09
1.5	93.74	71.18	81.31	87.45	78.07	84.24	76.34	82.26

Table 6. The performance when considering different loss functions on VeRi-776 datasets.

Methods	VeRi-776	
	Rank-1	mAP
Cross-Entropy	94.04	73.43
Circle	95.29	73.46
Circle + triplet	96.31	**78.36**
Cross-Entropy + triplet	**96.36**	78.32
Cross-Entropy + Contrast	94.93	72.93
Cross-Entropy + Lifted	95.35	74.43
Cross-Entropy + Instance	95.05	73.16

can be observed, both Rank-1 accuracy and mAP increase slowly as β rises. When β is set as 1, the model obtains the best performance. The performance declines as β continues to rise. This indicates that the joint learning of the two loss functions consistently brings more promising improvement by raising the value of β moderately.

The impact of different loss functions. To further explore the impact of different loss functions, we conduct several comparison experiments on VeRi-776 datasets. As illustrated in Table 6, compared with the cross-entropy loss, circle loss[55] achieves better performance on both mAP and Rank-1 accuracy. The result of utilizing the cross-entropy loss alone is lower than that of combining with a metric learning-based loss for example the instance loss[82], lifted loss[47], contrastive loss[15] or triplet loss. For our method, the circle+triplet method and the cross-entropy+triplet method obtain comparable results.

4.5 Complexity analysis

In Table 7, we report the number of parameters, FLOPs, inference time (seconds per image), and the corresponding performance for GB and SSR-Net. As shown in the results from Table 7, the addition of two extra components brings a small increase of 1.10×10^7(+31.88%) in the number of parameters required and 1.09×10^9(+11.85%) FLOPs computational cost while achieving 3.05% higher Rank-1 accuracy and 6.78% higher mAP. The inference time of the two methods is comparable, with only 4.6×10^{-4} seconds difference. Therefore, SSR-Net is computationally efficient as it requires a similar number of parameters while achieving better performance.

4.6 Comparison with the State-of-the-art

Evaluation on VehicleID. As illustrated in Table 8, we compare our proposed SSR-Net with several other methods on the VehicleID dataset. We categorize these methods into global feature extraction based methods,

Table 7. Complexity analysis of GB and SSR-Net.

Method	Parameters	FLOPs	Inference Time (sec/img)	VeRi-776	
				Rank-1	mAP
GB	2.35×10^7	8.11×10^9	2.27×10^{-3}	93.31	71.54
SSR-Net	3.45×10^7	9.20×10^9	2.73×10^{-3}	96.36	78.32

local feature extraction based methods and GCN-based methods which consider the relationship among local and global features. It can be observed that our method SSR-Net achieves a good performance when compared with other approaches. On the VehicleID-800 dataset, in comparison with approaches that only extract global features, our SSR-Net improves Rank-1 accuracy by at least 2.18% and mAP by at least 2.07%. On the middle-large VehicleID-1600 dataset, SSR-Net obtains 79.51% on Rank-1 and 85.13% on mAP respectively, which outperforms other methods by at least 0.71% in terms of Rank-1 and 0.93% in terms of mAP. SSR-Net obtains 77.44% on Rank-1 accuracy and 83.12% on mAP on the more difficult VehicleID-2400 dataset. Our method improves Rank-1 accuracy by at least 5.57% on Rank-1 and 7.77% on mAP when compared with GCN-based methods. The results from Table 8 demonstrate the superiority of our proposed SSR-Net, which jointly exploits the relationships among local regions and the relationships between local and global parts of the vehicles.

Evaluation on VeRi-776. We also validate our proposed SSR-Net on VeRi-776 dataset and compared the performance with several other approaches which are classified as global feature extraction, local feature extraction and GCN-based methods. The experimental results are summarized in Table 9. From Table 9 we can observe that the proposed method outperforms other approaches, with Rank-1 and mAP values of 96.36% and 78.32%, respectively. To be more specific, our proposed SSR-Net exceeds at least 2.42% in terms of mAP compared with methods that simply extract global features from image appearances. Compared with methods that consider extracting local features[16] and other additional information such as spatio-temporal information[54, 63, 86] and license plate appearance[18], the SSR-Net method outperforms them by at least 0.95% in higher Rank-1 accuracy and 0.24% in higher mAP performance. Compared with the GCN-based method, SSR-Net exceeds by at least 2.98% on Rank-1 and with 7.73% on mAP. These results illustrate the effectiveness and superiority of our proposed method.

4.7 Visualization of retrieval results

Image retrieval can be divided into two cases, single-gallery-shot and multi-gallery-shot. Note that we conduct experiments under the single-gallery-shot condition. Fig. 5 and Fig. 6 show several vehicle retrieval results on the VeRi-776 dataset in single-gallery-shot and multi-gallery-shot, respectively. Note that the leftmost image is the query image in each sub-figure, surrounded by a blue border, while images with green and red borders represent correct and wrong retrieval results, respectively. In Fig. 5 and Fig. 6, two rows of images in each sub-figure depict the retrieval results of the baseline and our proposed SSR-Net, respectively.

Single-gallery-shot refers to the case in which one sample per vehicle ID is chosen at random to construct a gallery of vehicles, while the other samples are utilized as query images. This implies that among the retrieval results list, there is just one sample image belonging to the same vehicle ID as the query image. The top-5 retrieval result lists of 6 query images are illustrated in Fig. 5. It can be seen that the proposed SSR-Net performs well for the vehicle re-identification task. In all cases, the SSR-Net method outperforms the baseline method according to the retrieved vehicle image results displayed in the first row of each sub-figure. In contrast, correct images retrieved by the baseline method do not rank in the first place in the retrieval results list. As it can be observed from Fig. 5 (f), the baseline does not even retrieve the correct sample in the top-5 ranked list, whereas the SSR-Net not only delivers the right result, but also ranks it at the top of the list.

Table 8. Performance comparison with other methods on VehicleID where * indicates that the same base network is used.

Methods	VehicleID-800		VehicleID-1600		VehicleID-2400	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
LOMO[33]	N/A	19.76	N/A	18.85	N/A	15.32
SCCN-CLBL-8 [84]	48.63	N/A	N/A	N/A	N/A	N/A
DRDL [35]	48.91	N/A	46.36	N/A	40.97	N/A
FACT [36]	49.53	N/A	44.63	N/A	39.91	N/A
ABLN[87]	52.63	N/A	N/A	N/A	N/A	N/A
FDA-Net [44]	N/A	N/A	59.84	65.33	55.53	61.84
SDC-CNN[89]	56.98	63.52	50.57	57.07	42.92	49.68
C2F-Rank [13]	61.10	63.50	56.20	60.00	51.40	53.00
Improved Triplet Loss[77]	69.90	N/A	66.20	N/A	63.20	N/A
GSTE [1]	75.90	75.40	74.80	74.30	74.00	72.40
BS[31]	78.80	86.19	73.41	81.69	69.33	78.16
UMTS [27]	80.90	87.00	78.80	84.20	76.10	82.80
VAMI [86]	63.12	N/A	52.87	N/A	47.34	N/A
TAMR [14]	66.02	N/A	62.90	N/A	59.69	N/A
QD-DLF[88]	72.32	76.54	70.66	74.63	64.14	68.41
OIFE+ST [63]	N/A	N/A	N/A	N/A	67.00	N/A
AAVER[28]	74.69	N/A	68.62	N/A	63.54	N/A
RAM [40]	75.20	N/A	72.30	N/A	67.70	N/A
EALN [45]	75.11	77.50	71.78	74.20	69.30	71.00
MAD+STR[26]	N/A	82.00	N/A	75.90	N/A	72.80
GRF+GGL[41]	77.10	N/A	72.70	N/A	70.00	N/A
PRN [3]	78.92	N/A	74.94	N/A	71.58	N/A
Part Regularization[16]	78.40	N/A	75.00	N/A	74.20	N/A
SAVER [29]	79.90	N/A	77.60	N/A	75.30	N/A
HSS-GCN*[69]	71.49	79.00	70.27	76.99	68.93	75.37
SGAT[90]	78.12	81.49	73.98	77.46	71.87	75.35
SSR-Net(Ours)	**83.08**	**89.07**	**79.51**	**85.13**	**77.44**	**83.12**

The multi-gallery-shot result illustration represents the situation where at least one image per vehicle ID constitutes the query image set, and the remaining sample images constitute the image gallery. The image gallery is made up of at least n samples per vehicle ID. In this experiment, n is set as 6, which implies that the top-10 retrieval result list only returns up to 6 vehicle images with the same ID as the query image. As depicted in Fig. 6, in comparison to the baseline approach, SSR-Net returns more correct samples despite the challenges in these examples posed by the diversity of changes in the view perspectives, while providing higher positive rankings in the image retrieval list. For example, in the upper row from Fig. 6 (b), we can observe that the baseline method

Table 9. Performance comparison with other methods on VeRi-776 where * indicates that the method is reproduced with the same base network.

Methods	Rank-1	mAP
LOMO[33]	25.30	9.60
SiameseVisual[54]	41.12	29.48
BOW-CN[81]	33.91	12.20
FACT [36]	52.21	18.75
GoogLeNet[73]	52.30	17.90
XVGAN[85]	60.30	24.70
ABLN[87]	60.49	24.92
SCCN-Ft+CLBL-8-Ft[84]	60.83	25.12
SDC-CNN[89]	83.49	53.45
FDA-Net [44]	84.27	55.49
GSTE [1]	N/A	59.47
VANet [5]	89.78	66.34
BS[31]	90.23	67.55
UMTS [27]	95.80	75.90
OIFE+ST [63]	92.40	51.42
Siamese-CNN+Path-LSTM[54]	83.49	58.27
QD-DLF [88]	88.50	61.83
VAMI+STR [86]	85.92	61.32
RAM [40]	88.60	61.50
EALN [45]	84.39	57.44
AAVER[28]	88.97	61.18
MAD+STR[26]	89.27	61.11
GRF+GGL[41]	89.40	61.70
PVSS[42]	90.58	62.62
PAMTRI[57]	92.86	71.88
Part Regularization [16]	94.30	74.30
SPAN[4]	94.00	68.90
Appearance+License[18]	95.41	78.08
HSS-GCN*[69]	93.38	70.59
SGAT [90]	89.69	65.66
SSR-Net(Ours)	**96.36**	**78.32**

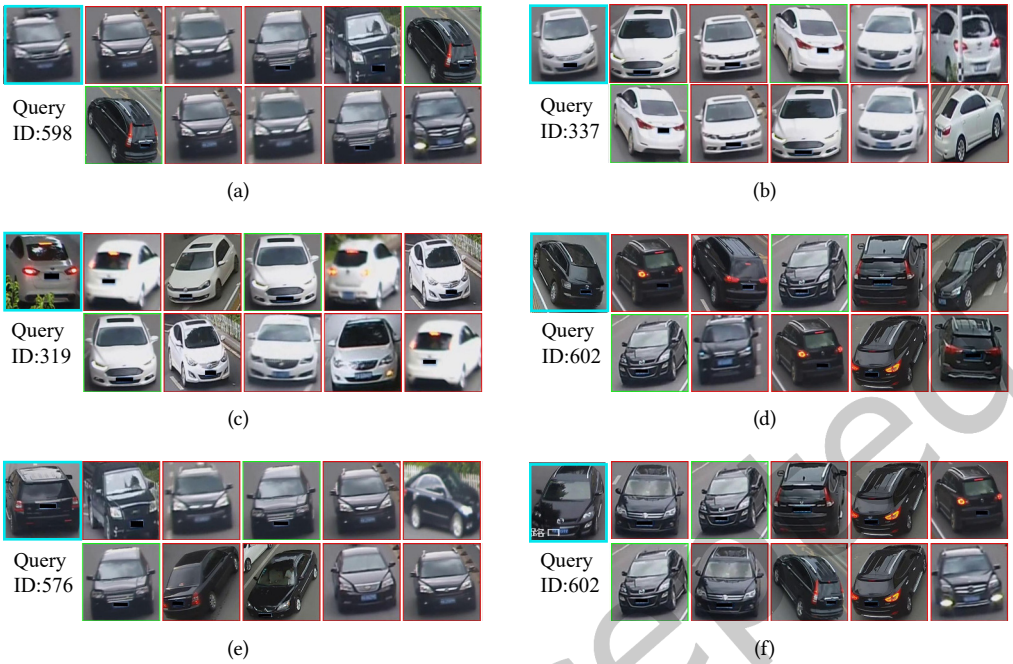


Fig. 5. Examples of Rank-5 retrieval results on VeRi-776 in single-gallery-shot condition. The queried vehicle image is shown in the left-most image. The top row in each example of queried vehicle image represents the results of the baseline while the bottom row illustrates the retrieved vehicle images by the proposed SSR-Net model.

retrieves only 4 correct images which are not even among the top choices. In contrast, SSR-Net not only that correctly retrieves 5 correct samples, but also sorts them appropriately in the top-5 positions of the list.

5 CONCLUSION

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62072027, Grant 61872032, and 62076021.

REFERENCES

[1] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. 2018. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia* 20, 9 (2018), 2385–2399.

[2] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203* (2013).

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. 2019. Partition and Reunion: A Two-Branch Neural Network for Vehicle Re-identification.. In *CVPR Workshops*. 184–192.

[4] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. 2020. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *European conference on computer vision*, Vol. LNCS 12347. Springer, 330–346.

[5] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. 2019. Vehicle re-identification with viewpoint-aware metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8282–8291.

[6] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1831–1840.

Fig. 6. Examples of Rank-10 retrieval results on VeRi-776 in multi-gallery-shot representations. The top row in each example of queried vehicle image represents the results of the baseline while the bottom row illustrates the retrieved vehicle images by the proposed SSR-Net model.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems* 29 (2016), 3844–3852.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[10] Gamaleldin F Elsayed, Simon Kornblith, and Quoc V Le. 2019. Saccader: Improving accuracy of hard attention models for vision. *arXiv preprint arXiv:1908.07644* (2019).
[11] Spyros Gidaris and Nikos Komodakis. 2019. Generating classification weights with gnn denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21–30.
[12] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 1263–1272.
[13] Haiyun Guo, Chaoyang Zhao, Zhiwei Liu, Jinqiao Wang, and Hanqing Lu. 2018. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[14] Haiyun Guo, Kuan Zhu, Ming Tang, and Jinqiao Wang. 2019. Two-level attention network with multi-grain ranking loss for vehicle re-identification. *IEEE Transactions on Image Processing* 28, 9 (2019), 4328–4338.
[15] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.
[16] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. 2019. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3997–4005.
[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[18] Yanguang He, Chenhe Dong, and Ying Wei. 2019. Combination of appearance and license plate features for vehicle re-identification. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3108–3112.
[19] Mikael Henaff, Joan Bruna, and Yann LeCun. 2015. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163* (2015).
[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
[21] Jinhui Hou, Huanqiang Zeng, Jianqing Zhu, Junhui Hou, Jing Chen, and Kai-Kuang Ma. 2019. Deep quadruplet appearance learning for vehicle re-identification. *IEEE Transactions on Vehicular Technology* 68, 9 (2019), 8512–8522.
[22] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. *Advances in neural information processing systems* 28 (2015).
[24] Haoxuanye Ji, Le Wang, Sanping Zhou, Wei Tang, Nanning Zheng, and Gang Hua. 2021. Meta Pairwise Relationship Distillation for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3661–3670.
[25] Bo Jiang, Xixi Wang, and Bin Luo. 2019. PH-GCN: Person re-identification with part-based hierarchical graph convolutional network. *arXiv preprint arXiv:1907.08822* (2019).
[26] Na Jiang, Yue Xu, Zhong Zhou, and Wei Wu. 2018. Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking. In *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 858–862.
[27] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2020. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11165–11172.
[28] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. 2019. A dual-path model with adaptive attention for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6132–6141.
[29] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. 2020. The devil is in the details: Self-supervised attention for vehicle re-identification. In *European Conference on Computer Vision*. Springer, 369–386.
[30] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[31] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. 2019. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.
[32] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).
[33] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.
[34] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*. 1449–1457.

[35] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. 2016. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2167–2175.
[36] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. 2016. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
[37] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. 2016. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *European conference on computer vision*. Springer, 869–884.
[38] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3566–3574.
[39] Xinchen Liu, Wu Liu, Jinkai Zheng, Chenggang Yan, and Tao Mei. 2020. Beyond the Parts: Learning Multi-view Cross-part Correlation for Vehicle Re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 907–915.
[40] Xiaobin Liu, Shiliang Zhang, Qingming Huang, and Wen Gao. 2018. Ram: a region-aware deep model for vehicle re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
[41] Xiaobin Liu, Shiliang Zhang, Xiaoyu Wang, Richang Hong, and Qi Tian. 2019. Group-group loss-based global-regional feature learning for vehicle re-identification. *IEEE Transactions on Image Processing* 29 (2019), 2638–2652.
[42] Xin-Chen Liu, Hua-Dong Ma, and Shuang-Qun Li. 2019. PVSS: A progressive vehicle search system for video surveillance networks. *Journal of Computer Science and Technology* 34, 3 (2019), 634–644.
[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030* (2021).
[44] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3235–3243.
[45] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. 2019. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing* 28, 8 (2019), 3794–3807.
[46] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
[47] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4004–4012.
[48] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. 2018. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514* (2018).
[49] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
[50] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.
[51] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
[52] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* (2015).
[53] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155* (2018).
[54] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. 2017. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proceedings of the IEEE International Conference on Computer Vision*. 1900–1909.
[55] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6398–6407.
[56] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*. 480–496.
[57] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. 2019. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 211–220.
[58] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. 2009. Learning color names for real-world applications. *IEEE Transactions on Image Processing* 18, 7 (2009), 1512–1523.
[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[60] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[61] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6857–6866.
[62] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* 38, 5 (2019), 1–12.
[63] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. 2017. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 379–387.
[64] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.
[65] Minghu Wu, Yeqiang Qian, Chunxiang Wang, and Ming Yang. 2021. A Multi-Camera Vehicle Tracking System Based on City-Scale Vehicle Re-ID and Spatial-Temporal Information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 4077–4086.
[66] Shijie Wu, Pamela Shapiro, and Ryan Cotterell. 2018. Hard non-monotonic attention for character-level transduction. *arXiv preprint arXiv:1808.10024* (2018).
[67] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 842–850.
[68] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.
[69] Zheming Xu, Lili Wei, Congyan Lang, Songhe Feng, Tao Wang, and Adrian Gheorghe Bors. 2021. HSS-GCN: A Hierarchical Spatial Structural Graph Convolutional Network for Vehicle Re-identification. In *Proc. ICPR's Int. Workshop on Human and Vehicle Analysis for Intelligent Urban Computing (IUC)*, Vol. LNCS 12665. Springer, 356–364.
[70] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4394–4402.
[71] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.
[72] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. 2020. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3289–3299.
[73] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3973–3981.
[74] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 974–983.
[75] Xiaotong Zhang, Han Liu, Qimai Li, and Xiao-Ming Wu. 2019. Attributed graph clustering via adaptive graph convolution. *arXiv preprint arXiv:1906.01210* (2019).
[76] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 1, 1 (2010), 43–52.
[77] Yiheng Zhang, Dong Liu, and Zheng-Jun Zha. 2017. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1386–1391.
[78] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. 2017. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia* 19, 6 (2017), 1245–1256.
[79] Jianan Zhao, Fengliang Qi, Guangyu Ren, and Lin Xu. 2021. PhD Learning: Learning With Pompeiu-Hausdorff Distances for Video-Based Vehicle Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2225–2235.
[80] Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. 2021. Heterogeneous Relational Complement for Vehicle Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 205–214.
[81] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.
[82] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.
[83] Zhedong Zheng, Liang Zheng, and Yi Yang. 2018. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2018), 3037–3045.
[84] Yi Zhou, Li Liu, and Ling Shao. 2018. Vehicle re-identification by deep hidden multi-view inference. *IEEE Transactions on Image Processing* 27, 7 (2018), 3275–3287.

[85] Yi Zhou and Ling Shao. 2017. Cross-View GAN Based Vehicle Generation for Re-identification.. In *BMVC*, Vol. 1. 1–12.
[86] Yi Zhou and Ling Shao. 2018. Aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6489–6498.
[87] Yi Zhou and Ling Shao. 2018. Vehicle re-identification by adversarial bi-directional lstm network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 653–662.
[88] Jianqing Zhu, Huanqiang Zeng, Jingchang Huang, Shengcai Liao, Zhen Lei, Canhui Cai, and Lixin Zheng. 2019. Vehicle re-identification using quadruple directional deep learning features. *IEEE Transactions on Intelligent Transportation Systems* 21, 1 (2019), 410–420.
[89] Jianqing Zhu, Huanqiang Zeng, Zhen Lei, Shengcai Liao, Lixin Zheng, and Canhui Cai. 2018. A shortly and densely connected convolutional neural network for vehicle re-identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 3285–3290.
[90] Yangchun Zhu, Zheng-Jun Zha, Tianzhu Zhang, Jiawei Liu, and Jiebo Luo. 2020. A Structured Graph Attention Network for Vehicle Re-Identification. In *Proceedings of the 28th ACM international conference on Multimedia*. 646–654.