Generic Representation Learning for Vehicle Association Guided by Foundational Models

Qilei Li, Mingliang Gao, *Senior Member, IEEE*, Jinyong Chen, Wenzhe Zhai, Gwanggil Jeon, *Senior Member, IEEE*, and Ahmed M. Abdelmoniem, *Senior Member, IEEE*

***Abstract*— Vehicle association is a vital yet complex task to retrieve specific vehicles across various camera angles, time frames, and geographical locations. In environments supported by autonomous driving and 6G networks, this task plays a vital role in urban surveillance and traffic management by enabling the real-time sharing of vehicle location and status information through ultra-high-speed, low-latency 6G communication. The success of a retrieval model largely depends on the quality of the extracted representations, which can be influenced by factors such as background diversity and occlusions. This study proposes a method to extract representations that remain consistent across different domains while retaining the discriminative power necessary to determine a vehicle's spatial location, regardless of background or environmental variations. To achieve this, we introduce a framework called Generic Representation Learning (GRL). Within GRL, we leverage large-scale pre-trained foundational models to provide spatial priors of vehicles, specifically the Grounding DINO model for object detection and the SAM model for object segmentation. These modules collaborate to help the network understand the spatial context of the object, enabling the feature extractor to focus on discriminative areas while minimizing interference. Additionally, we introduce a complementary feature alignment mechanism based on a memory bank to explore globally applicable knowledge within the learned representation of the object. These constituent elements collectively form SRP, to enhance its capability for outstanding performance in vehicle retrieval. Extensive experimentation demonstrates that SRP significantly outperforms existing models on widely recognized benchmarks.**

***Index Terms*— Vehicle association, road planning, deep neural network, object segmentation and location.**

I. Introduction

VEHICLE association has emerged as a pivotal research area with the advancements in artificial intelligence of thing (AIoT) and plays a critical role in autonomous driving and 6G networks. It focused on retrieving vehicles from images or video sequences captured by multiple non-overlapping surveillance cameras [1]. With the ultra-low latency and high-speed communication provided by 6G networks, vehicle association not only significantly improves the efficiency of traffic systems but also provides smarter solutions for urban planning. Particularly in the field of autonomous driving, the high reliability and large bandwidth of 6G enable real-time data transmission and response between vehicles and infrastructure. This capability enhances the perception accuracy, decision-making speed, and path planning precision of autonomous vehicles, thereby optimizing the management and operational efficiency of the entire traffic system. Fig. 1 illustrates the application in autonomous driving systems [2], [3].

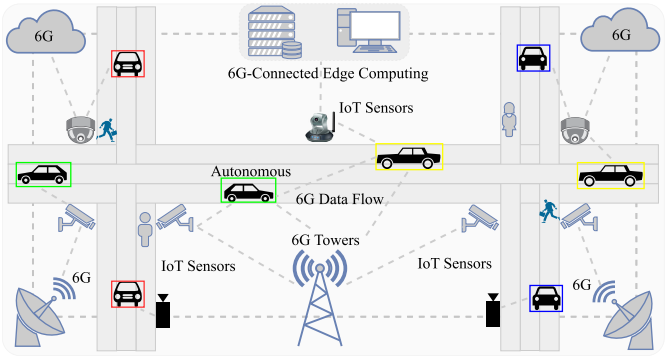


Fig. 1. IoT-based vehicle road planning system with real-time monitoring and analysis capabilities, leveraging 6G-enabled vehicle-to-infrastructure communication to enhance autonomous driving performance, optimize traffic flow, and improve urban planning efficiency.

In recent years, smart cities and smart cars have been popularized and applied [4]. The development of the vehicle detection system prompts the rapid improvement of the vehicle association model. A vehicle detection model can provide a bounding box for each detected vehicle object, as shown in Fig. 2 (a). Vehicle association is a subsequent process of identifying vehicles and tracking their motion trajectories from cropped vehicle images. The application of a practical vehicle association system is shown in Fig. 2 (b), which has received much attention. Unlike general classification tasks, vehicle association presents unique challenges, such as scale and attitude variations, background noise, and differentiation between vehicles with similar appearances.

Received 22 August 2024; revised 30 October 2024 and 1 January 2025; accepted 23 February 2025. The Associate Editor for this article was I. Ashraf. *(Corresponding author: Mingliang Gao.)*

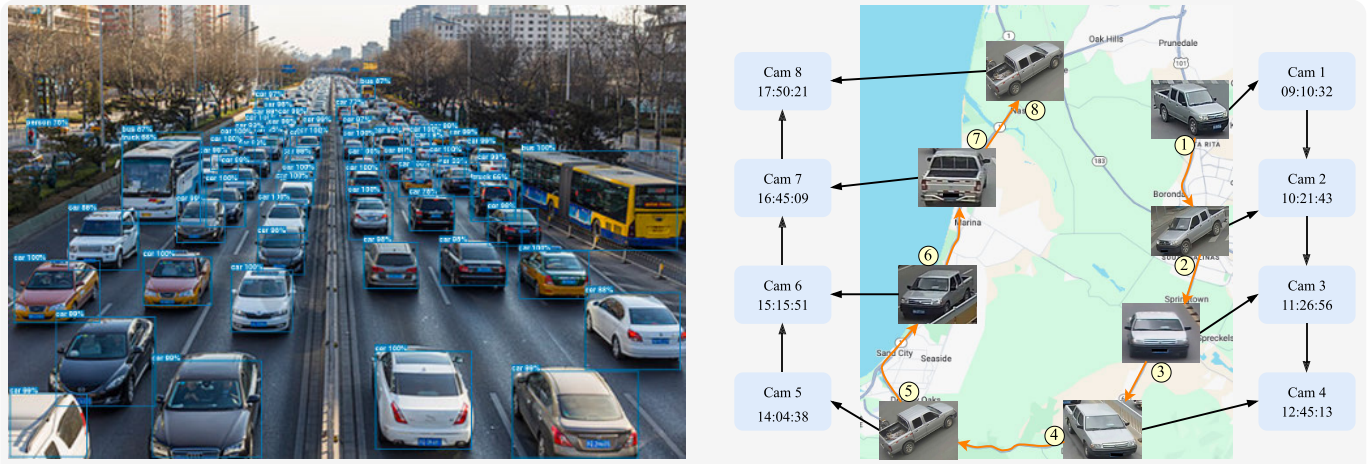
Qilei Li is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: q.li@qmul.ac.uk).

Mingliang Gao, Jinyong Chen, and Wenzhe Zhai are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: mlgao@sdut.edu.cn; jinyongch@outlook.com; wenzhezhai@outlook.com).

Gwanggil Jeon is with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Ahmed M. Abdelmoniem is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: ahmed.sayed@qmul.ac.uk).

Digital Object Identifier 10.1109/TITS.2025.3545910



(a) Vehicle detection with bounding box (b) Application of Vehicle Association System

Fig. 2. Visual Results of a vehicle detection model and vehicle association system.

Early methods [5], [6] in vehicle association primarily relied on convolutional neural networks (CNNs) to acquire global image features while overlooking local features, which limited performance. Huang et al. [7] introduced a dense convolutional network that enhances the propagation of features while reducing the number of parameters. Subsequent advancements [8], [9] integrated CNNs for global feature learning and introduced a graph network (GN) branch to explore relationships among local features. Liu et al. [10] proposed PCRNet to learn more discriminative local features and explore correlations between local regions. However, the convolution and downsampling operations at certain depths hindered association performance [11]. The lack of compatibility between CNNs and GNs, constrained by separate loss functions, further limited the association's performance. To tackle these challenges, the Transformer methods [12], [13], [14] have gained traction in vehicle association. It utilizes a multi-head self-attention module to capture global information and establish long-distance dependencies, eliminating downsampling operations in conventional CNN methods. However, the Transformer method often entails high computational costs and intricate network architectures. A Semantic-Oriented Feature Coupling Transformer (SOFCT) [15] network is designed to automate semantic features and amplify feature expression. However, these methods often ignore the extraction of spatial information, which hinders the extraction of image context information and improved association performance. Furthermore, many of these methods are trained and validated independently on a common dataset. This practice, marked by the uniformity of the entire background and image style within the dataset, somewhat diverges from real-world scenarios. This hinders the improvement of the model's generalization ability. This poses a challenge for vehicle associations.

In this work, we propose Generic Representation Learning (GRL) to address the challenges encountered by existing methods in vehicle association. The primary objective is to enhance the feature extractor's awareness of the spatial location of the vehicle object. This is achieved by leveraging spatial priors generated by large-scale pre-trained models. These models, trained on diverse data, can ground the spatial region despite shifts in data distribution. Specifically, we utilize Grounding DINO for vehicle localization, which is a calibration to rectify inaccurate bounding boxes in the input image. Additionally, we employ the Segment Anything Model (SAM) to delineate the fine-grained spatial regions of the object. These two models collaboratively provide hierarchical priors that aid in making the multi-scale representation spatially aware, thereby mitigating distribution shifts caused by various factors. Moreover, we introduce a memory-bank-based representation alignment module that aligns local representations with a global memory. This alignment ensures that representations converge into a unified feature space and share semantic features, facilitating distance metric measurements. The contributions of this paper are three-fold:

- We introduce the idea of guiding discriminative regional-aware representation learning by external prior guiding. In that regard, we explore Grounding DINO for vehicle localization, which is a calibration to rectify inaccurate bounding boxes in the input image.
- We include the Segment Anything Model (SAM) to delineate the fine-grained spatial regions of the object, providing hierarchical priors that aid in making the multi-scale representation spatially aware.
- We design a memory-bank-based representation alignment module that aligns local representations with a global memory, ensuring convergence into a unified feature space and facilitating distance metric measurements.

The remainder of the paper is structured as follows: In Section II, we introduce the literature related to the proposed method. In Section III, the methodology of the proposed model is detailed. In Section IV, Comprehensive experiments and ablation studies are conducted to validate the efficiency of the proposed method. The conclusion is drawn in Section V.

II. Related Works

A. Vehicle Association

Object association is an important research task within computer vision that recognizes the same objects across

different camera views, spanning both videos and images captured by surveillance cameras. Vehicle association is a significant sub-topic in association that aims to identify the same vehicle across different images. This technology is essential in numerous applications, such as drone monitoring and autonomous driving. Vehicle association technology enables explainable intelligence augmented for vehicular road cooperation. Accurately identifying and tracking the same vehicle at different times and locations provides essential data support and interpretability for decision-making in intelligent transportation systems [3]. Vehicle association involves identifying vehicles by extracting global and local features and auxiliary information, such as color and brand. In recent years, vehicle associations have garnered significant attention from researchers, driven by the rapid advancement of deep learning technology. Shen et al. [8] employs convolutional neural networks of different depths to extract global features. Moreover, the existing approaches often fail to capture the local information. Consequently, several studies [16], [17] have focused on extracting local features to complement global features. Typically, these approaches divide the feature map into distinct regions, encompassing horizontal, vertical, and circular areas. For example, Xu et al. [18] divided the feature graph into five regions and utilized the graph convolutional network (GCN) to represent the spatial relationships among the features. However, these methods frequently face challenges due to limited generalization capabilities. To address this issue, Meng et al. [19] proposed an analytic network that divides the vehicle into four views, thereby enhancing the accuracy of local feature extraction. In addition, the attention mechanism has also attracted interest in vehicle association. Lee et al. [20] uses multiple soft attention points to soften the vehicle feature map to identify vehicles. Zhu et al. [21] proposed a dual attention module to learn distinct regional dependencies and enhance spatial awareness of local features. Shen et al. [22] combined GCN and Transformers to extract global and local features and capture their interaction and correlation. Numerous researchers have proposed viable methods to tackle the issue of vehicle association. In this work, we aim to learn a lightweight but advanced vehicle association model, that can serve as a strong backbone for vehicle data analysis and traffic planning.

B. Vehicle Road Cooperation in AIoT

Traffic planning plays a pivotal role in the sustainable development of city [23]. Hence, fortifying the establishment and administration of urban traffic planning is fundamental for fostering cities' sound and systematic development. The advancements in vehicle association technology enable more accurate surveillance and regulation of vehicle maneuvers within urban road networks, furnishing crucial technical support for urban traffic planning [24]. On the one hand, vehicle association technology is crucial for real-time traffic management and control. During peak traffic periods or special events, real-time monitoring of vehicle movements enables timely adjustments to traffic signals, lane control strategies, and other measures to alleviate congestion and optimize traffic flow [25], [26]. On the other hand, vehicle association technology enhances vehicle-road collaboration, thereby improving road usage efficiency. Analyzing vehicle driving data yields insights into road usage patterns and driver behaviors to form a scientific foundation for road design [27], [28]. Furthermore, vehicle association technology can be employed for anomaly detection to enhance traffic safety [29]. This information provides more precise data support for traffic planning, assisting planners in optimizing road layouts and traffic signal controls, thereby boosting the efficiency and capacity of the traffic system. Consequently, integrating urban traffic planning and vehicle association technology is crucial for traffic planning. In summary, the integration of vehicle association technology with urban traffic planning provides critical tools for real-time traffic management, road optimization, and anomaly detection. This technology offers valuable insights into vehicle behaviors and road usage, leading to more informed and efficient urban planning decisions. By leveraging vehicle association data, urban planners can design better road layouts, improve traffic control strategies, and ultimately contribute to the sustainable development of cities.

C. Spatial Prior Mining With Foundation Models

Traditional saliency detection tasks typically depend on well-designed models. Saliency detection aims to identify the most visually prominent regions in an image and provide a coarse understanding of spatial relationships within the image. Saliency detection models localize and classify objects within an image, offering detailed information about their locations in the scene. These tasks are fundamental to computer vision and have been addressed by various techniques and algorithms. Traditional methods typically process the information from a single modality, thereby lacking the ability to leverage supplementary information from other modalities. Recent advancements in foundation models present more powerful alternatives for spatial reasoning [30], [31]. For instance, Segment Anything Model (SAM) [31] empowers models to focus on specific regions of an image based on their relevance to the task. This enables a more dynamic and data-driven approach to spatial reasoning. Grounding DINO [30] is a recently released zero-sample target detector that achieves good results on several datasets. It combines the Transformer-based detector DINO with truthful pre-training, which can introduce linguistic information into target detection to enable the recognition of new categories. These foundation models have found widespread application in various domains, including Medical Image Segmentation [32], Remote Sensing [33], and 3D Object Segmentation [34]. For example, Chen et al. [35] utilized the semantic segmentation map output from SAM to guide the visual coherence learning of foreground and background features. This approach effectively solves the problem of ignoring neighboring priors in traditional global-level feature matching. Osco et al. [36] highlighted the utility of SAM for processing aerial and orbital images from diverse geographical contexts. In this paper, we introduce the idea of learning invariant representation by the assistance of spatial priors, extracted with large-scale pretrained foundation models.

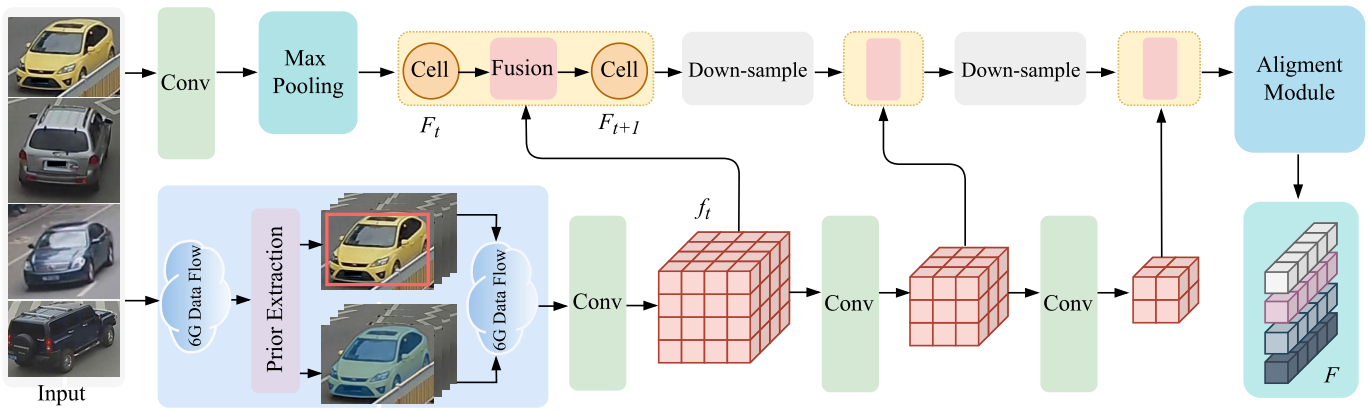


Fig. 3. Architecture of the Generic Representation Learning (GRL) model for vehicle association. The GRL model consists of three key modules: a feature extraction model (upper part) for hierarchical features learning, a prior extraction module (lower part) to calibrate the spatial priors, and a representation alignment module (tail) for fine-tuning vehicle representations with the memory-bank-based prototype.

D. Domain-Invariant Knowledge Learning

In order to obtain discriminative representation for vehicle association, it's essential that the feature extractor can focus on domain-invariant knowledge. The task of domain invariant representation learning [37] presents significant challenges due to the diverse characteristics of data across different domains. This poses a major difficulty in learning invariant representations, particularly in tasks like object association, where images captured in different domains exhibit differences in background complexity [38] and occlusion [39]. One solution is to locate objects accurately, yet this remains a significant challenge in vehicle association. To address this issue, some existing works have attempted to address background issues in object association by employing techniques such as background subtraction [40], [41], [42] or scene parsing [43], [44]. Sajid and Cheung [41] employed a flexible pixel classification method capable of accurately distinguishing foreground from background. Simultaneously, multiple color spaces handled color information and better differentiated between various objects and backgrounds. Meng et al. [43] proposed a perspective alignment method, which maps vehicle images into 3D space better to align the features of vehicles from different perspectives. These models handle and analyze both global and local features within vehicle images, while also considering the spatial information of the vehicles. However, despite these efforts, the domain generalization problem remains particularly challenging due to inconsistent data distributions [45], [46]. The aforementioned methods often struggle to generalize effectively across different domains, which leads to insufficient robustness and performance degradation when faced with unseen data distributions [37]. One of the primary obstacles in the association is the issue of domain-invariant knowledge generalization [47]. Some existing works have focused on domain-invariant learning methods, such as Domain Adaptation Label Smoothing Regularization [48] and Adversarial Target Invariant Representation Learning [49], which can assist the model in better adapting to different data distributions. These methods aim to improve the generalization ability of unseen domains. Additionally, Ni et al. [50] designed a proxy task named Cross-ID Similarity Learning (CSL) to mine shared local visual information among different data. CSL enables the model to learn universal features by focusing only on partial visual similarity, thus alleviating the side effects of domain-specific biases. Zhao et al. [51] proposed the Memory-based Multi-Source Meta-Learning framework, which simulates the training-testing process of domain generalization by introducing meta-learning strategies, thereby learning a more generalizable model. In this work, we overcome the problem of data distribution shift within vehicle data captured in different conditions and proposed to learn domain-invariant representation with the additional guidance of spatial priors.

III. Methodology

A. Overall Framework

In this section, we provide a detailed introduction of the Generic Representation Learning (GRL) framework, crafted to cultivate discriminative representations tailored for vehicle association. Depicted in Fig. 3, the GRL model comprises three foundational modules: (1) A feature extraction backbone is proposed to process vehicle images and extract hierarchical features across various scales; (2) A prior extraction fusion model that effectively integrates spatial priors obtained from input images into intermediate feature representations and (3) A representation alignment module designed to fine-tune the vehicle representation utilizing a memory-bank based generic prototype. This adjustment facilitates the creation of a uniform feature space encompassing all samples. These three modules, constructed as adaptable units, can be amalgamated into a unified framework facilitating end-to-end training via conventional gradient updates. Subsequent sections delve into the intricate design aspects of these modules. The notations are summarized in Table I.

B. Hierarchical Feature Extraction

This experiment employs vehicle or human images as the source image for the network. Drawing inspiration from MSINet [52], we exploit the advantages of lightweight models

TABLE I
MATHEMATICAL NOTATIONS

Notation	Description
F	Input image
$\text{Cel}(\cdot)$	Network module in each block
$\text{Conv}(\cdot)$	Convolution operation
$\text{Concat}(\cdot)$	Concatenation operator
$\text{Proj.}(\cdot)$	Non-local projection operator
$\text{EMA}(\cdot)$	Exponential moving average
W_{global}	Global calibration process
W_{local}	Local calibration process

to address challenges related to multi-scale and spatial information. The specific network structure diagram is illustrated in Fig. 3. The MSINet [52] network architecture is employed as the underlying structure for feature extraction. Moreover, drawing inspiration from the methodology of OSNet [53], images undergo processing via a 7×7 convolutional layer and a max-pooling layer to bolster the network's capability in extracting sophisticated features. Following this, feature processing is divided into three stages, and each stage is composed of two cells and a fusion module. In each stage, a series of two cascaded cells are interconnected, with the initial cell altering the channel count. The output of the initial cell is then channelled into the fusion module along with the spatial prior by the fusion module, which can integrate guiding knowledge to guide the representation being attentive to the vehicle's spatial information. Subsequently, the second cell takes in the fused representation for further feature refinement. The formulation is represented as follows:

$$F'_{t+1} = \text{Cell}_b(\text{Fusion}(\text{Cell}_a(F_t), f_t)), \quad t \in \{0, 1, 2\}, \tag{1}$$

where F_t is the input of the stage t, F_{t+1} is the output of the stage, and f_t is the prior knowledge. Each stage is composed by two $\text{Cell}(\cdot)$ blocks, which is the feature extraction block that was proposed in MSINet [52]. After the first stage, a downsampling operator decreases the width and height of the feature dimension by half.

$$F_t = \text{AvgPool}(\text{Conv}(F'_t)), \quad t \in \{1, 2\}. \tag{2}$$

In Equation (2), the Conv operator employs a 1×1 kernel to reduce dimensions, with a stride of 2 applied for spatial pooling. Following the downsampling operation, the features undergo additional processing in the subsequent stages. It's worth noting that the downsampling operator is applied for the first stages. The output of the third stage, namely F_3 is used as the final representation to conduct distance matching during testing.

C. Spatial Prior Remedy

The core of the proposed GRL framework is incorporating spatial information about objects into the acquired intermediate representation. This spatial data is an external prior capable of directing the model's attention toward the object region. With the advancement of deep learning methods, we extract prior information using two state-of-the-art vision models, *i.e.,* the Grounding DINO model [30], employed for object detection, and the Segment Anything Model (SAM) [31] for object semantic segmentation. These models excel in identifying object locations and offer extensive capabilities for expansion, courtesy of their multi-modal prompt design, enabling precise object specification through textual instructions. However, directly incorporating prior information as additional input proves suboptimal due to its inherent modality gap with the raw image. Although retraining the prior extraction network presents a potential solution, the associated training process requires substantial computing resources and also demands a significant volume of training data to ensure convergence. Alternatively, drawing inspiration from recent Adaptor designs in Large-Language Model research literature, we propose freezing the prior extractor and designing a lightweight cascaded convolutional network (CNN) to establish a connection between prior knowledge and intermediate feature representation. This process involves three primary steps of the prior, namely feature extraction, adaptation, and fusion.

1) Prior Extraction: The term prior denotes external knowledge that assists in data understanding. The understanding of spatial distribution is crucial in vehicle association. Hence, we employ advanced object grounding and segmentation techniques for prior information extraction. We classify priors into two distinct types with a coarse-to-fine strategy: Grounding DINO for object detection and the Segment Anything Model (SAM) for semantic segmentation. Both models are adaptable and can be guided by textual instructions, such as 'vehicle' in the context of vehicle detection. Given the vehicle detection task, an input image I undergoes processing within the prior extraction network. This network preprocesses the input image to ensure alignment with its requirements. Subsequently, the preprocessed data is sent through the prior extraction modules, which are structured as a parallel fusion of Grounding DINO and SAM. These modules operate concurrently to produce distinct prior knowledge outputs. Specifically, Grounding DINO outputs coordinate information $b = (x_1, y_1, x_2, y_2)$ representing bounding boxes, and SAM outputs a binary mask M that indicates the region of the object.

2) Prior Adaptation: The two priors provided by Grounding DINO and SAM convey analogous knowledge but operate on different granularity levels and dimensions. To consolidate them into a unified module, we employ a bounding box approach on a zero image, *i.e.,* a black image as a background, with dimensions matching the input image's dimensions. This method essentially entails annotating a mask at the box level, denoted as B. To handle these two priors effectively, we concatenate them along the channel dimension and input them into a lightweight cascaded CNN for hierarchical prior refinement. These layers function as adaptors, aligning the priors from their respective original spaces with the hyperspace of the representations derived from the main branch. To process these two priors, at each step, the concatenated prior is iteratively optimized. This process is formulated as follows:

$$\begin{aligned} P_0 &= \text{Concat}(B, M), \\ \hat{P}_{t+1} &= \text{Adaptor}_t(P_t), \quad \text{s.t.} \quad t \in \{0, 1, 2\}, \end{aligned} \tag{3}$$

where P_t represents the transformed prior at step t. The Adaptor module is structured according to the conventional "ReLU-BN-Conv" design [54].

3) Prior Fusion: Once the per-stage prior is adapted, it becomes incorporated into the vehicle representation to enrich its features. This integration is accomplished through a fusion module, which includes a concatenation operator and a CNN layer with a 1×1 kernel for aggregating knowledge and reducing channels. This process is formulated as

$$\hat{F}_t = \text{Conv}_t(\text{Concat}(P_t, F_t)), \quad \text{s.t.} \quad t \in \{1, 2, 3\}. \tag{4}$$

D. Memory-Bank Based Representation Calibration

Since association is a subsequent step to object detection, where the bounding box of the target object is provided, it is reasonable to assume that the object occupies a comparable spatial position. Motivated by this premise, we propose a representation calibration module based on memory banks to facilitate spatial alignment. Specifically, for a given batch of representations F_3', we propose a local and global calibration module to maintain local and global memory banks as reference points. The two memory banks are initially set randomly and then updated using the Exponential Moving Average (EMA) with the current representations from the batch. Regarding global calibration, non-local projections are employed to create two counterparts: Key and Query. Subsequently, the Key updates the global memory bank and interacts with the Query to compute the per-sample activation score. The global calibration process is formulated as follows:

$$W_{\text{global}} = \text{Proj}_Q(F_3') \times \text{EMA}(\text{Mean}(\text{Proj}_K(F_3'))), \tag{5}$$

where $\text{Proj}_{\cdot}(\cdot)$ indicates the non-local projection. Correspondingly, the raw representation is obtained from the CNN layer to update the local memory bank for the local calibration. An activation matrix is computed to encode the local variation between the raw representation and the global memory bank by multiplying the sampled partial memory. The local calibration procedure is structured as follows:

$$W_{\text{local}} = F_3' \times \mathbb{1}(\text{EMA}(\text{Mean}(F_3')), \tag{6}$$

where $\mathbb{1}(\cdot)$ denotes the local random batch sampling process. To utilize local and global activations, we employ the position activation module (PAM) [52] to achieve alignment of the positive and negative samples in the current batch simultaneously. The formula is formulated as,

$$\mathcal{L}_{\text{align}} = \text{PAM}(W_{\text{global}}, W_{\text{local}}), \tag{7}$$

where $\mathcal{L}_{\text{align}}$ indicates the align loss of the calibration module.

E. Learning Objective

A blend of learning objectives guides the proposed method: comprising cross-entropy loss, triplet loss, and the newly introduced calibration loss in Eq.(7). The triplet loss is calculated as follows:

$$\mathcal{L}_{\text{tri}} = \left[D(f_a, f_p) - D(f_a, f_n) + \omega)\right]_+, \tag{8}$$

where $D(f_a, f_p)$ represents the Euclidean distance between positive features for the anchor, while $D(f_a, f_n)$ represents the Euclidean distance between negative features for the anchor. The notation 'a', 'n', and 'p' refers to anchor, negative, and positive samples used in the triplet loss. ω is the margin, and $[\cdot]_+$ represents the $max(\cdot, 0)$ function. The cross-entropy loss is calculated as follows:

$$\mathcal{L}_{\text{cross}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} t_{i,j} log(p_{i,j}), \tag{9}$$

where N denotes the batch size, C represents the number of classes, $t_{i,j}$ stands for the value of the target label j for sample i, and $p_{i,j}$ signifies the prediction probability of the model for the sample class.

Summing up, the final loss function is formulated by combining the three individual objectives:

$$\mathcal{L}_{\text{all}} = \alpha\mathcal{L}_{\text{tri}} + \beta\mathcal{L}_{\text{cross}} + \gamma\mathcal{L}_{\text{align}}, \tag{10}$$

The α, β and γ are the hyper-parameters to balance individual items.

IV. Experiments

A. Implementation Details

The experiments were conducted following two protocols: in-domain and cross-domain experiments. The learning rate was set to 0.065 for the in-domain experiment, and the margin parameter was set to 0.3. The model was trained for 350 epochs. The cross-domain experiments are designed to assess the model's generalization. To avoid model over-fitting to the source domain, it was trained for 250 epochs with the other hyper-parameters, the same as in the in-domain experiments. Throughout the training process, the weight loss of the alignment module was set to 2.0, ensuring its integration with the triplet loss and cross-entropy loss as components of the loss function. The proposed model was implemented on PyTorch, with all experiments conducted on a server equipped with an NVIDIA A100 GPU. The hyper-parameters α, β, γ in Eq. (10) were set to 1, 1, and 10, respectively.

B. Datasets and Evaluation Metrics

1) Datasets: We examined the efficacy of the proposed GRL model on two extensive vehicle datasets, VeRi-776 [55] and VehicleID [56]. The VeRi-776 dataset comprises over 50,000 images depicting 776 vehicles captured by 20 cameras from various viewing angles and occlusion scenarios. The training set includes 37,778 images from 576 vehicles, while the test set contains 11,579 images from the remaining 200 vehicles. The VehicleID dataset [56] is a large-scale urban dataset focusing on vehicle association, with 221,763 images portraying 26,267 cars of approximately 250 vehicle types with color variations. Additionally, to verify the versatility of the proposed framework across different object counting tasks, we also evaluated it on two person association datasets, namely Market-1501 [57] and MSMT17 [58].

2) Evaluation Metrics: This work uses the mean Average Precision (mAP) to measure the average precision of search results across all queried image IDs.

$$\text{mAP} = \frac{\sum_m^M \text{AP}(m)}{M}. \tag{11}$$

where M denotes the total number of images, and $\text{AP}(m)$ represents the average precision of the m-th image. The

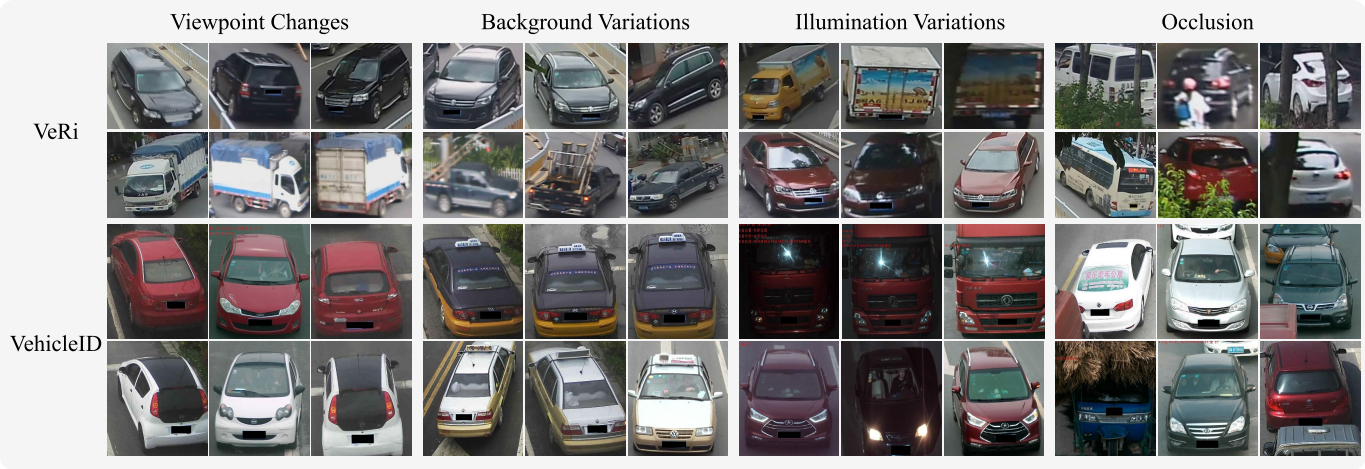


Fig. 4. Illustration of within/cross-domain data distribution shift caused by various factors, such as viewpoint changes, background variations, illuminations variations, and occlusion.

TABLE II
COMPARISON OF RE-IDENTIFICATION (RE-ID) PERFORMANCE ON VARIOUS DATASETS. (THE BEST RESULTS ARE HIGHLIGHTED IN RED)

Methods	#Params.	VeRi		VehicleID		Market1501		MSMT17	
		R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	mAP↑	R-1↑	R-5↑
ResNet50 [59]	∼24M	92.8	69.9	70.6	76.6	85.7	68.3	48.0	25.7
OSNet [53]	2.2M	95.4	72.8	76.0	88.7	93.6	81.0	71.0	43.3
CDNet [60]	1.8M	94.3	73.0	74.5	88.8	93.7	83.7	73.7	48.5
MSINet [52]	2.3M	95.9	75.0	76.5	**89.8**	94.6	87.0	76.0	52.5
GRL (Ours)	2.3M	**95.9**	**79.0**	**78.6**	84.5	**95.5**	**89.7**	**82.6**	**61.8**

Cumulative Matching Characteristic (CMC) curve is a holistic assessment metric for association performance. It entails sorting the similarity between queried images and those in the dataset, subsequently determining the probability that the top-k retrieved images include the correct query results. The CMC score for Rank-k is commonly computed by summing the maximum value of each query image and dividing it by the total number of query images.

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

To evaluate the effectiveness of the proposed network, we conducted a comprehensive analysis covering intra-domain and cross-domain person association scenarios. These comparisons verified the generalizability of the proposed method.

The experimental results are depicted in Table II. Noteworthy is the superior performance of GRL over the other methods across most metrics. Although the proposed method did not achieve the top mAP on the VehicleID dataset, it demonstrated the highest Rank-1 accuracy across all datasets and achieved the best mAP on the VeRi-776 [55], Market-1501 [57], and MSMT17 [58] datasets. This discrepancy could be due to the VehicleID dataset's limited sample diversity, which lacks a mix of positive and negative match examples. As a result, the model may be less sensitive to negative matches, despite its capacity to extract discriminative representation for accurate positive matching. To mitigate this issue, one potential solution is to employ hard example mining so that such contrastive learning can help the model not only achieve top-ranked results but also become more sensitive to negative matches.

TABLE III
COMPARISON ON PERSON REID BENCHMARKS AGAINST SOTA METHODS. (THE BEST RESULTS ARE HIGHLIGHTED IN RED)

Method	Market-1501		MSMT17	
	Rank1↑	mAP↑	Rank1↑	mAP↑
PCB [61]	93.8	81.6	68.2	40.4
MGN [62]	95.7	86.9	76.9	52.1
OSNet [53]	93.6	81.0	71.0	43.3
IANet [63]	94.4	83.1	75.5	46.8
DGNet [64]	94.8	86.0	77.2	52.3
Auto-ReID [65]	94.5	85.1	-	-
CDNet [60]	95.1	86.0	78.9	54.7
BAT-Net [66]	95.1	87.4	79.5	56.8
SFT [67]	94.1	87.5	79.0	58.3
CTF [68]	94.8	87.7	-	-
MSINet [52]	95.3	89.6	81.0	59.6
GRL (Ours)	**95.6**	**89.8**	**82.4**	**62.1**

To validate the generalization performance of proposed model, Table III presents a comparison between the proposed method and other state-of-the-art (SOTA) methods on Market-1501 [57] and MSMT17 [58]. Compared with BAT-Net [66], the GRL achieves a 0.5% improvement in Rank-1 accuracy on the Market1501 dataset, increasing from 95.1% to 95.6%, and a 2.4% improvement in mAP, increasing from 87.4% to 89.8%. Additionally, compared with MSINet [52] on the MSMT17 [58] dataset, the proposed method improves the

Fig. 5. Visualization of input sample, the generated mask from SAM, and bounding box from Grounding DINO model.

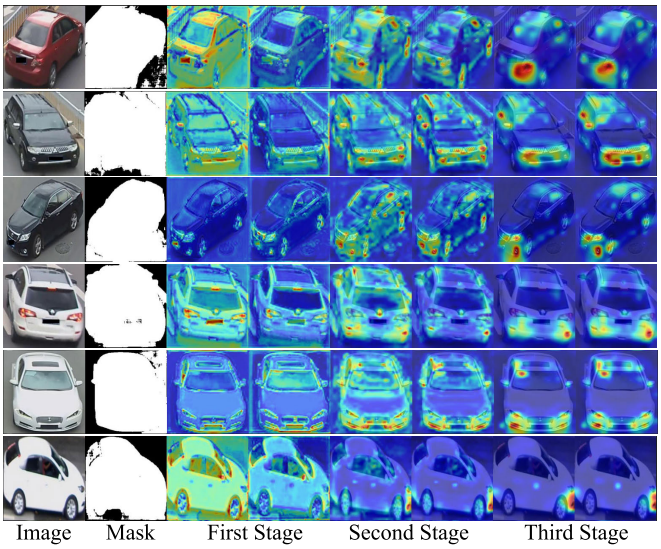


Fig. 6. Visualization of the query vehicle, the segmentation mask, intermediate feature activations from the baseline (left), and proposed GRL (right).

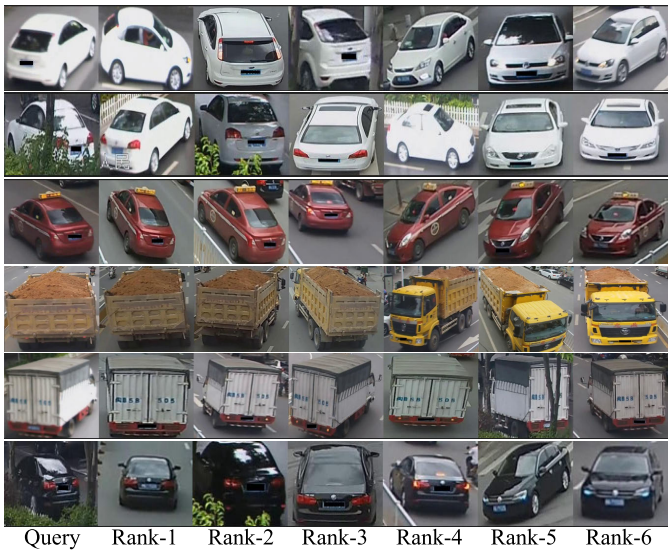


Fig. 7. Visualization of the query vehicle and the corresponding two matches.

Rank-1 accuracy by 1.4%, from 81.0% to 82.4%, and the mAP by 2.5%, from 59.6% to 62.1%. These results demonstrate the significant performance enhancement of the proposed method across different datasets, which validates its excellent generalization capability.

D. Visualization Analysis

1) Visualization on Spatial Prior: We visualized the extracted spatial priors to validate the reliability of the prior extraction module in extracting spatial priors. The details are shown in the Fig. 5. Spatial priors facilitate the model to attentively focus on the object by excluding disruptive background interference. The bounding box and mask precisely identify the target object in the image.

2) Visualization of Hierarchical Feature Representations: We performed a sequence of visualizations to evaluate the features' characteristics preceding and succeeding the application of the fusion module throughout the three stages. Our model pays more attention to the object itself and minimizes extraneous background noise, as demonstrated in Fig. 6. The proposed GRL method can effectively extract the spatial prior and focus on the target object, which is crucial for achieving robust and accurate vehicle association. Specifically, the use of masks as priors offers more accurate and informative guidance, which makes the feature extraction process more effective. Moreover, the blocks in different stages extract features that evolve from semantic (concrete) to abstract, which allows the network to progressively refine its feature representations. Finally, with the guidance of prior information, the GRL model can focus better on the main target (*i.e.,* the vehicle) in the early stages and gradually attend to more discriminative regions, such as the front and rear of the vehicle in later stages. These factors together improve the overall performance of the model and validate the effectiveness of the proposed model.

3) Visualization on Top Match Results: In addition, we visualized the top matching results in Fig. 7. It can be observed that the proposed method can accurately retrieve the query image from the gallery, regardless of the background and environmental variations.

V. Conclusion

In this work, we proposed a Domain-Invariant Representation Learning framework with Generic Representation Learning (GRL) to achieve robust and accurate vehicle association across various surveillance scenarios. By leveraging foundational models such as the Grounding DINO for object detection and the SAM model for object segmentation, GRL

significantly improved the spatial context comprehension of the network. This approach enabled the feature extractor to focus more on discriminative regions of the vehicles, effectively reducing the influence of background noise and occlusions. Additionally, integrating a memory-bank-based feature alignment mechanism allowed for the incorporation of globally relevant knowledge into the learned representations. These innovations collectively enhanced the performance of the vehicle association system, enabling it to surpass existing models by a significant margin on established benchmarks. Ultimately, this study addressed the key challenges in vehicle association and demonstrated the potential of tailored, domain-invariant representation learning strategies to advance city surveillance and traffic management systems.

REFERENCES

[1] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *Comput. Vis. Image Understand.*, vol. 182, pp. 50–63, May 2019.

[2] H. Wang, J. Hou, and N. Chen, "A survey of vehicle re-identification based on deep learning," *IEEE Access*, vol. 7, pp. 172443–172469, 2019.

[3] A. Amiri, A. Kaya, and A. S. Keceli, "A comprehensive survey on deep-learning-based vehicle re-identification: Models, data sets and challenges," 2024, *arXiv:2401.10643*.

[4] X. Chen, H. Yu, C. Hu, and H. Wang, "Multi-branch feature learning network via global-local self-distillation for vehicle re-identification," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12415–12425, Sep. 2024.

[5] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1900–1909.

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

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, Jun. 2017, pp. 4700–4708.

[8] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8793–8804, Jul. 2022.

[9] Z. Xu, L. Wei, C. Lang, S. Feng, T. Wang, and A. G. Bors, "HSS-GCN: A hierarchical spatial structural graph convolutional network for vehicle re-identification," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2021, pp. 356–364.

[10] X. Liu, W. Liu, J. Zheng, C. Yan, and T. Mei, "Beyond the parts: Learning multi-view cross-part correlation for vehicle re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 907–915.

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

[12] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[13] B. Yu et al., "High-performance discriminative tracking with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9856–9865.

[14] B. Kim, J. Lee, J. Kang, E. Kim, and H. J. Kim, "HOTR: End-to-end human-object interaction detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 74–83.

[15] Z. Yu, Z. Huang, J. Pei, L. Tahsin, and D. Sun, "Semantic-oriented feature coupling transformer for vehicle re-identification in intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2803–2813, Mar. 2024.

[16] X. Chen, H. Yu, F. Zhao, Y. Hu, and Z. Li, "Global–local discriminative representation learning network for viewpoint-aware vehicle re-identification in intelligent transportation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[17] L. Liang, C. Lang, Z. Li, J. Zhao, T. Wang, and S. Feng, "Seeing crucial parts: Vehicle model verification via a discriminative representation model," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 1s, pp. 1–22, Feb. 2022.

[18] Z. Xu et al., "SSR-Net: A spatial structural relation network for vehicle re-identification," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 6, pp. 1–22, Dec. 2022.

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

[20] S. Lee, T. Woo, and S. H. Lee, "Multi-attention-based soft partition network for vehicle re-identification," *J. Comput. Des. Eng.*, vol. 10, no. 2, pp. 488–502, Mar. 2023.

[21] W. Zhu et al., "A dual self-attention mechanism for vehicle re-identification," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109258.

[22] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng, "GiT: Graph interactive transformer for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 1039–1051, 2023.

[23] J. Rui and F. Othengrafen, "Examining the role of innovative streets in enhancing urban mobility and livability for sustainable urban transition: A review," *Sustainability*, vol. 15, no. 7, p. 5709, Mar. 2023.

[24] Zakria et al., "Trends in vehicle re-identification past, present, and future: A comprehensive review," *Mathematics*, vol. 9, no. 24, p. 3162, Dec. 2021.

[25] J. Tu, C. Chen, X. Huang, J. He, and X. Guan, "Discriminative feature representation with spatio-temporal cues for vehicle re-identification," 2020, *arXiv:2011.06852*.

[26] C. Sun, Y. Wang, Y. Deng, H. Li, and J. Guo, "Research on vehicle re-identification for vehicle road collaboration," *J. Phys., Conf. Ser.*, vol. 2456, no. 1, Mar. 2023, Art. no. 012025.

[27] Z. Xiong, M. Li, Y. Ma, and X. Wu, "Vehicle re-identification with image processing and car-following model using multiple surveillance cameras from urban arterials," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7619–7630, Dec. 2021.

[28] B. A. Holla, M. M. M. Pai, U. Verma, and R. M. Pai, "Enhanced vehicle re-identification for smart city applications using zone specific surveillance," *IEEE Access*, vol. 11, pp. 29234–29249, 2023.

[29] K.-T. Nguyen et al., "Vehicle re-identification with learned representation and spatial verification and abnormality detection with multi-adaptive vehicle detectors for traffic video analysis," in *Proc. CVPR Workshops*, Jan. 2019, pp. 363–372.

[30] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," 2023, *arXiv:2303.05499*.

[31] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[32] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," 2023, *arXiv:2304.13785*.

[33] D. Wang, J. Zhang, B. Du, D. Tao, and L. Zhang, "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 8815–8827.

[34] J. Cen et al., "Segment anything in 3D with NeRFs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 25971–25990.

[35] H. Chen, Y. Li, Z. Gu, Z. Xu, J. Lan, and H. Li, "Segment anything model meets image harmonization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2024, pp. 2630–2634.

[36] L. P. Osco et al., "The segment anything model (SAM) for remote sensing applications: From zero to one shot," 2023, *arXiv:2306.16623*.

[37] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2022.

[38] R. Wei, J. Gu, S. He, and W. Jiang, "Transformer-based domain-specific representation for unsupervised domain adaptive vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 2935–2946, Mar. 2023.

[39] Y. Bai, J. Liu, Y. Lou, C. Wang, and L.-Y. Duan, "Disentangled feature learning network and a comprehensive benchmark for vehicle re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6854–6871, Oct. 2022.

[40] S. Agrawal and D. P. Natu, "Segmentation of moving objects using numerous background subtraction methods for surveillance applications," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 3, pp. 2553–2563, Jan. 2020.

[41] H. Sajid and S. S. Cheung, "Universal multimode background subtraction," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3249–3260, Jul. 2017.

[42] S. K. Choudhury, P. K. Sa, S. Bakshi, and B. Majhi, "An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios," *IEEE Access*, vol. 4, pp. 6133–6150, 2016.

[43] D. Meng, L. Li, X. Liu, L. Gao, and Q. Huang, "Viewpoint alignment and discriminative parts enhancement in 3D space for vehicle ReID," *IEEE Trans. Multimedia*, vol. 25, pp. 2954–2965, 2022.

[44] D. Meng, L. Li, S. Wang, X. Gao, Z.-J. Zha, and Q. Huang, "Fine-grained feature alignment with part perspective transformation for vehicle ReID," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 619–627.

[45] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit, "On calibration and out-of-domain generalization," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 2215–2227, 2021.

[46] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 719–728.

[47] Y. Ding, L. Wang, B. Liang, S. Liang, Y. Wang, and F. Chen, "Domain generalization by learning and removing domain-specific features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 24226–24239.

[48] Q. Wang et al., "Inter-domain adaptation label for data augmentation in vehicle re-identification," *IEEE Trans. Multimedia*, vol. 24, pp. 1031–1041, 2022.

[49] G. Gupta, R. Kapila, K. Gupta, and R. Raskar, "Domain generalization in robust invariant representation," 2023, *arXiv:2304.03431*.

[50] H. Ni, Y. Li, L. Gao, H. T. Shen, and J. Song, "Part-aware transformer for generalizable person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11280–11289.

[51] Y. Zhao et al., "Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6273–6282.

[52] J. Gu et al., "MSINet: Twins contrastive search of multi-scale interaction for object ReID," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 19243–19253.

[53] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.

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

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

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

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

[58] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[59] H. Luo, W. Jiang, Y. Gu, F. Liu, and X. Liao, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2019.

[60] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 6729–6738.

[61] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.

[62] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[63] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.

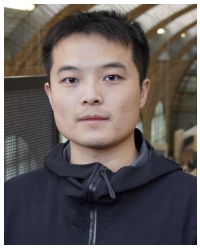
[64] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2133–2142.

[65] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-ReID: Searching for a part-aware ConvNet for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3750–3759.

[66] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8030–8039.

[67] C. Luo, Y. Chen, N. Wang, and Z.-X. Zhang, "Spectral feature transformation for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4976–4985.

[68] A. Zhang, Y. Gao, Y. Niu, W. Liu, and Y. Zhou, "Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 598–607.



Qilei Li received the M.S. degree from Sichuan University in 2020 and the Ph.D. degree in computer science from the Queen Mary University of London. From June 2022 to April 2024, he worked as a Machine Learning Scientist with Veritone Inc., where he focused on developing a scalable person search framework for retrieving individuals at different locations and times, as captured by various cameras. His current research interests include privacy-aware distributed machine learning, with a particular emphasis on learning domain-invariant knowledge representation from multimodal data captured in diverse environments. His research outcome has been recognized as an ESI Highly Cited Paper (Top 1%). Additionally, he serves as an Evaluator for the ELLIS PhD Program.

Mingliang Gao (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from Sichuan University. He was a Visiting Lecturer with The University of British Columbia from 2018 to 2019. He is currently an Associate Professor and the Vice Dean of the School of Electrical and Electronic Engineering, Shandong University of Technology. He has been the Principal Investigator for a variety of research funding, including the National Natural Science Foundation, China Postdoctoral Foundation, and the National Key Research Development Project. He has published over 200 journal/conference papers in IEEE, Springer, Elsevier, and Wiley. His research interests include computer vision, machine learning, and intelligent optimal control. He is an Associate Editor of *Expert Systems* and *Network Modeling Analysis in Health Informatics and Bioinformatics*.

Jinyong Chen is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include information fusion, multimodal object counting, and deep learning.

Wenzhe Zhai is currently pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo. His research interests include smart city systems, information fusion, crowd analysis, and deep learning. Additionally, he serves as a reviewer for numerous journals, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Neurocomputing*, EAAI, and *Multimedia Systems*.

Gwanggil Jeon (Senior Member, IEEE) received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, South Korea, in 2003, 2005, and 2008, respectively. From September 2009 to August 2011, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. From September 2011 to February 2012, he was with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, as an Assistant Professor. From December 2014 to February 2015 and from June 2015 to July 2015, he was a Visiting Scholar with the Centre de Mathématiques et de Leurs Applications (CMLA), École Normale Supérieure Paris-Saclay (ENS-Cachan), France. From 2019 to 2020, he was a Prestigious Visiting Professor with the Dipartimento di Informatica, Università degli Studi di Milano Statale, Italy. He was a Visiting Professor with Sichuan University, China; the Universitat Pompeu Fabra, Barcelona, Spain; Xinjiang University, China; the King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand; and the University of Burgundy, Dijon, France. He is currently a Full Professor with Incheon National University, Incheon, South Korea. He was a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and the Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020. He is an Associate Editor of *Sustainable Cities and Society*, IEEE ACCESS, *Real-Time Image Processing*, *Journal of System Architecture*, and *MDPI Remote Sensing*.

Ahmed M. Abdelmoniem (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2017. Formerly, he was a Research Scientist with KAUST, Saudi Arabia, and a Senior Researcher with Huawei's Future Networks Laboratory, Hong Kong. He is currently an Associate Professor with the School of Electronic Engineering and Computer Science, Queen Mary University of London, U.K. He is an investigator on projects totaling USD 1.5mil in funding. His work appears in top-tier conferences and journals, including NeurIPS, AAAI, MLSys, ACM EuroSys, IEEE INFOCOM and ICDCS, IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and *Computer Networks* (Elsevier). His research interests include the intersection of distributed systems, networks, and machine learning.