IEEE Open Journal of
Intelligent Transportation
Systems

Received 4 September 2024; revised 30 December 2024; accepted 26 January 2025. Date of publication 3 February 2025; date of current version 3 March 2025.

Digital Object Identifier 10.1109/OJITS.2025.3538037

Vehicle Re-Identification and Tracking: Algorithmic Approach, Challenges and Future Directions

ASHUTOSH HOLLA B. [1], MANOHARA M. M. PAI [2] (Senior Member, IEEE), UJJWAL VERMA [3] (Senior Member, IEEE), AND RADHIKA M. PAI [1] (Senior Member, IEEE)

[1]Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
[2]Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
[3]Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India

CORRESPONDING AUTHORS: M. M. M. PAI, U. VERMA, AND R. M. PAI (e-mail: mmm.pai@manipal.edu; ujjwal.verma@manipal.edu; radhika.pai@manipal.edu)

This article has supplementary downloadable material available at https://doi.org/10.1109/OJITS.2025.3538037, provided by the authors.

ABSTRACT Vehicle re-identification and tracking play a vital role in intelligent transportation systems as they enhance traffic management, improve safety, and optimize flow by precisely monitoring and analyzing vehicle movements across various locations. This technology enables the collecting of data in real-time, which allows for effective identification of incidents, enforcement of laws, and decision-making in urban planning. Deep learning techniques used in vehicle re-identification extract distinct characteristics to identify and match a vehicle across different camera perspectives. This bridges the non-overlapping field of camera views and forms a relationship between the detected vehicles. Tracking enhances this process by assigning a distinct identifier to the recognized vehicle, allowing for the creation of a continuous trajectory across the network for further analysis. Vehicle re-identification and tracking have made substantial progress in recent years as a result of the accelerated development of deep learning. Consequently, it is imperative to conduct a thorough examination of these chores. To provide a detailed picture of the research towards vehicle re-identification and tracking, this study provides the recent advancements of various datasets, and frameworks and strategies undertaken to perform these tasks. Specifically, the paper provides a comprehensive review of the different modes of re-identification of vehicles and further analysis. The paper also discusses the challenges and directions that can be taken in future for vehicle re-identification and tracking.

INDEX TERMS Deep learning, intelligent transportation systems, vehicle re-identification, vehicle tracking.

I. INTRODUCTION

TO ENHANCE the overall quality of life, Smart Cities provide people living there with unparalleled revolutionary services that include improved service quality, reduced cost, reduced environmental impact, and substantially increased comfort. A critical challenge in daily transportation within smart cities is the growing number of fatal road accidents. Factors such as irresponsible driving, distracted driving, driving under the influence of drugs or alcohol, and poor road conditions can be attributed to this increase, which results in physical, emotional, and financial damage to those concerned. Ensuring road safety is vital in smart cities due to the growing urbanization of modern urban surroundings [1]. Utilizing cutting-edge technologies is crucial for reducing risks and establishing safer societal systems. With the help of ITS, the present transportation system may operate more effectively, safely, and comfortably while reducing negative environmental effects. To monitor and control traffic flow, ITS employ a range of surveillance technologies, including cameras, sensors, GPS data, and crowd sourcing information via applications. Predictive analytics and traffic modeling are made possible by the real-time processing of this data and its storage for later examination. Adaptive responses

The review of this article was arranged by Associate Editor Jianwu Fang.

and pattern recognition are further improved by AI and machine learning. Applications of this data include dynamic message signs, adaptive traffic signals, real-time incident detection, and traveler information systems. Taken together, these technologies optimize traffic flow, increase safety, lower emissions, and offer insightful information for policy and infrastructure planning. Surveillance data is essential for AI and computer vision applications in traffic management, enabling functionalities such as vehicle detection [2], [3], [4], [5], counting [6], [7], re-identification [8], [9], and tracking [5], [10].

Vehicle re-identification is of utmost importance in intelligent transportation systems (ITS). Conventional traffic monitoring methods depend on data obtained from a single camera, which restricts its capacity to monitor vehicles over long distances or at crossings where the camera's field of view does not overlap. Vehicle re-identification in Intelligent Transportation Systems (ITS) enables the ongoing surveillance of specific cars across several camera perspectives in extensive traffic networks. Using deep learning algorithms, it examines characteristics such as the color, manufacture, and model of a vehicle to determine its distinct identification. By utilizing this capability, ITS successfully identify, and link identical vehicles captured by different camera angles, therefore efficiently connecting the missing information between them. Furthermore, vehicle re-identification enables sophisticated functionalities such as vehicle theft prevention and investigation, as it enables law enforcement to track the movement of suspicious vehicles across many places.

For ITS to continuously monitor a vehicle's trajectory inside a camera's field of view or across a network of cameras, vehicle tracking is essential. It is necessary for real-time traffic management, which makes it possible to anticipate and reduce traffic, improves road safety by utilizing collision detection and avoidance technologies, and aids law enforcement in observing and handling infractions. Vehicle re-identification and vehicle tracking have different approaches and scopes, but they both aim to monitor vehicles. Vehicle tracking employs temporal consistency to track a moving vehicle in a continuous sequence inside the field of view of a single camera. For vehicle Re-ID, on the other hand, vehicles must be recognized and matched across non-overlapping camera views. To manage these variations in perspective, illumination, and partial occlusions, strong feature extraction and matching algorithms are required. ITS extensively utilize surveillance cameras strategically positioned to localize and track vehicles across a network of cameras. These cameras provide critical data inputs for deep learning models, which analyze vehicle features for accurate re-identification. However, CCTV-based re-identification faces significant challenges: (1) Limited field of view can hinder the capture of comprehensive vehicle information needed for re-identification, (2) Expanding CCTV infrastructure is cost-prohibitive and labor-intensive, and (3) Variations in brightness, illumination changes, and occlusions can drastically reduce re-identification accuracy [11], [12], [13]. Recently, the increased use of UAVs has shown promise in overcoming these challenges due to their mobility, ability to capture high-resolution data, wide field of view, and dynamic altitude capabilities. Consequently, researchers have focused on leveraging UAVs for vehicle re-identification, leading to significant advancements in the field.

Initially, the process of vehicle re-identification was dependent on sensors [14], [15], [16] that estimated the travel of individual vehicles by matching vehicle signatures detected at various locations. This method was subsequently improved by vision-based methods, which focused on the identification of vehicles by utilizing hand-crafted features such as edges, corners, and textures. Furthermore, certain techniques utilized license plate recognition from surveillance cameras to re-identify vehicles [17], [18], capitalizing on the distinctive identifiers of license plates. Nevertheless, license plate information, which is uniquely identifiable, raises privacy concerns, despite its potential to enhance re-identification accuracy [19]. This method becomes even more complex in multilingual countries by the inconsistencies in license plate characters that result from noncompliance with regulations. Vehicle re-identification has made significant strides, particularly with the proliferation of deep learning techniques [20], [21], because of the emergence of advanced storage solutions and edge computing technologies [22], [23]. The extensive research that has been conducted to address the remaining challenges in vehicle re-identification has been motivated by the success of Convolutional Neural Networks (CNNs) in a variety of computer vision tasks. The accuracy and robustness of vehicle re-identification systems can be significantly enhanced by the automatic extraction and learning of complex features from large datasets by these deep learning models. This change has resulted in an increase in the number of publications and research endeavors that are designed to overcome the obstacles that conventional methods were unable to effectively address.

In recent years, the application of transformers has been extended to vision-related tasks as a result of the increasing popularity of transformers in natural language processing (NLP) [24], [25], [26] for a variety of text analysis tasks. This has been achieved through the development of a variety of transformer architectures. These architectures are intended to achieve performance that is equivalent to or superior to that of conventional Convolutional Neural Networks (CNNs) in a variety of image and video-based vision tasks. Methods, datasets, and metric learning adaptations have been the subject of numerous comprehensive surveys [27], [28] that have addressed vehicle re-identification and tracking. The vehicle re-identification survey presented by [27] discussed vehicle re-identification studies that are categorized into sensor based and vision-based techniques (Figure 1). Their study mainly focused on different sensor-based methods for re-identification by emphasizing techniques such as magnetic sensors, GPS RFID and cellular phones, hybrid

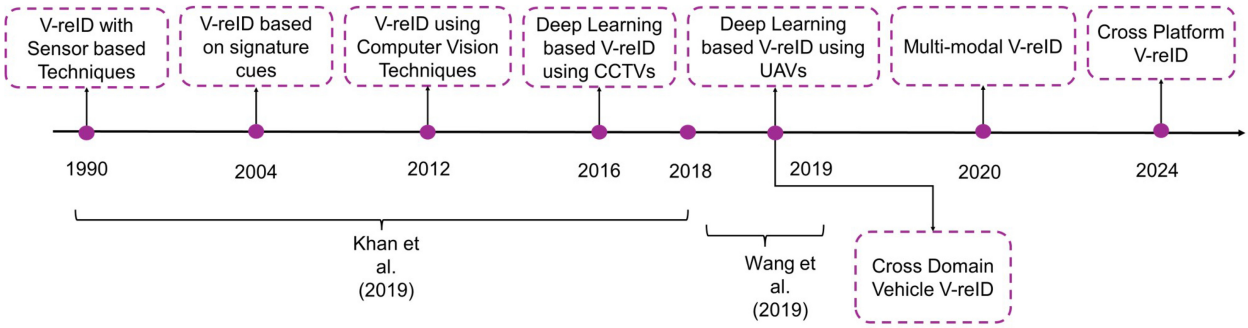

FIGURE 1. A timeline of the development of vehicle re-identification methods over the years.

methods etc. Contributions for vehicle re-identification using deep learning techniques was further surveyed by Wang et al. [28] (Figure 1). Their study on re-identification survey focused on vision-based approaches, metric learning based vehicle re-identification using deep features, attention mechanism, etc. The work also highlights the datasets, challenges and future directions for performing vehicle re-identification. Nevertheless, most of this work is limited to 2019. Few studies presented in [27], [28] are still used as a benchmark dataset/framework to evaluate the developed re-identification framework. Post 2019 several forms of vehicle re-identification were designed and introduced by various studies. This paper offers a comprehensive examination of the most recent developments in vehicle re-identification, with a particular emphasis on the most recent transformer and CNN architectures. It investigates a variety of re-identification methods, such as cross domain, multi-modality, and cross platform vehicle re-identification. Furthermore, the paper emphasizes the substantial research contributions that were made in the AI City Challenge with respect to vehicle re-identification and tracking.

This study presents recent advancements in vehicle re-identification and tracking, detailing various approaches to address the challenges in this field. Based on the taxonomy outlined in Figure 2, the study bridges research gaps by summarizing prominent existing works on re-identification. The survey primarily focuses on different forms of re-identification using surveillance cameras, multi-spectral inputs, techniques for adapting vehicles observed in domains different from the source domain, and the use of cross platform surveillance cameras, such as CCTVs and UAVs. The study also underscores the significance of vehicle tracking by highlighting recent advancements in CCTV and UAV surveillance videos. The survey encompasses research contributions made post-2019, offering insights into the most relevant and promising developments in vehicle re-identification and tracking.

The rest of the paper is organized as follows. Section II presents the survey on the work for four different forms of vehicle re-identification techniques and the metrics considered for evaluating the re-identification algorithms. Section III discusses the significant contribution of vehicle tracking that are developed for CCTV or UAV videos. Section IV briefs about different datasets that have been developed for various forms of re-identification and also for tracking. Section V discusses the challenges and future pathways for vehicle re-identification and tracking. Finally the study concludes the work in Section VI.

II. MODES OF VEHICLE RE-IDENTIFICATION

A. SINGLE PLATFORM VEHICLE RE-IDENTIFICATION

Vehicle re-identification is an important task in intelligent transportation systems. Its goal is to identify a specific vehicle across different surveillance systems that are used in areas of high activity. These systems, typically consisting of CCTVs and UAVs, record the movement of traffic from different perspectives. Researchers utilize this abundant data to address the difficult task of identifying vehicles across these diverse platforms. This study explores recent advancements in deep learning for vehicle re-identification, in contrast to previous efforts that primarily focused on individual platforms using traditional machine learning or emerging deep learning techniques. More precisely, we investigate the incorporation of attention mechanisms, innovative re-identification loss metrics, and transformer architectures into these frameworks. This analysis sheds light on how these cutting-edge deep learning techniques are pushing the boundaries of vehicle re-identification accuracy and robustness.

The vehicle re-identification process using deep learning typically involves a feature extractor designed to capture the semantic attributes of vehicles. Commonly used feature extractors include standard architectures like ResNets [29], DenseNet [30], and transformer models such as ViT [31]. Some research studies incorporate attention mechanisms to address specific re-identification challenges by enriching the datasets with additional annotations. These networks are trained using publicly available datasets that provide annotations for identical vehicles observed across multiple surveillance cameras. The deep learning-based re-identification algorithms are trained using either supervised, unsupervised, or self-supervised approaches. Figure 3 demonstrates a general re-identification training method using supervised learning, employing categorical cross-entropy and triplet loss. Regardless of the re-identification approach, the training process involves optimizing the feature

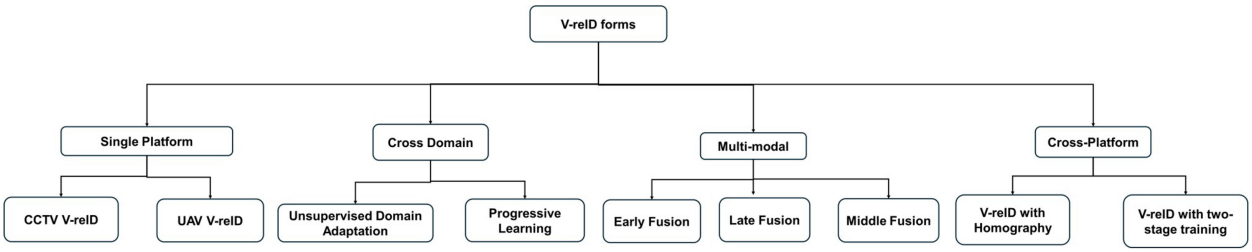

FIGURE 2. An illustration of the taxonomy of various forms in vehicle re-identification. Vehicle Re-identification modes are divided into Single Platform, Cross Domain, Multi-modal, and cross platform. Subcategories are provided for each mode that includes CCTV and UAV V-reID for Single Platform, Progressive Learning and Unsupervised Domain Adaptation for Cross Domain, and fusion techniques for Multi-modal. Furthermore, recent advancement in vehicle re-identification with cross platform mode contains categories of re-identification with Homography and Two-Stage Training.

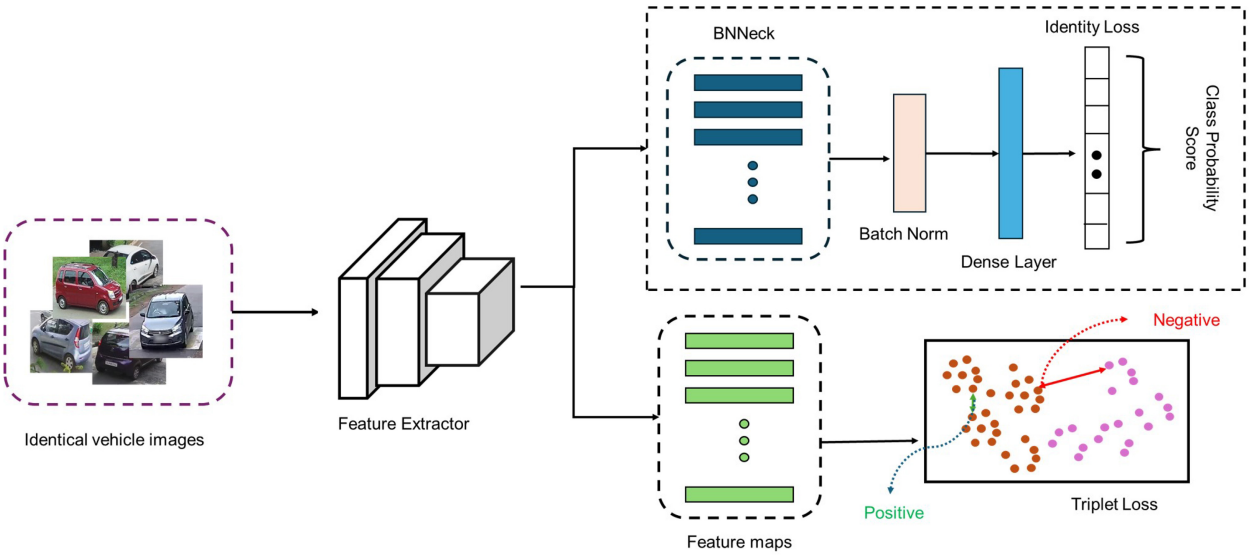

FIGURE 3. Illustration of different phases in a vehicle re-identification architecture: A feature extractor processes identical vehicle images to produce feature maps. These feature maps are generally fed to two branches, i.e., (a) a classification branch with Bottleneck, batch normalisation, and a dense layer to compute class probabilities using identity loss. (b) Metric learning branch using triplet loss to bring positive samples closer in the embedding space while separating negatives.

extractor using various loss functions, such as pairwise loss, triplet loss, and contrastive loss, to enhance the model's ability to distinguish between different vehicles.

Contrastive learning has emerged as an effective self-supervised learning paradigm, notably influential in domains such as re-identification, where the ability to differentiate between fine-grained differences between instances (vehicle/person) is essential. Contrastive learning fundamentally promotes the proximity of representations of similar data points (positives) in the embedding space, while simultaneously displacing those of dissimilar data points (negatives). This is accomplished by utilizing contrastive loss functions, such as Triplet Loss [32] or InfoNCE [33], which utilize paired information to direct the training of the embedding network [34]. Contrastive learning frequently necessitates fewer labeled samples than conventional supervised methods, as it relies on augmentations or inherent structures within the data to determine similarity. At the initial research towards re-identification the contrastive learning techniques were primarily designed for person entity. The techniques such as cross-input neighborhood differences [35], a proxy classifier learning with a co-training strategy [36] and a Siamese network architecture [37] to learn the discriminative feature embeddings incorporate contrastive learning.

Towards the task of vehicle re-identification, a novel multi-branch network called CFVMNet is proposed by the authors of [38] for vehicle re-identification. The purpose of the study was to address re-identification challenges such as minor inter-class differences and orientation variations. The CFVMNet extracts global and local detail features using four branches and a Batch DropBlock (BDB) [39] strategy to enhance inter-class differences. Additionally, vehicle attributes like color, type, and model were considered to improve feature recognition. For the issue of orientation variation leading to large intra-class differences, the CFVMNet learns two metrics based on whether there is a common field of view, allowing it to focus on different regions. The CFVMNet architecture involves two-level features corresponding to common and different fields of view, split into four branches. Authors have experimented on two datasets, i.e., VeRi-776 [40] and VehicleID [8] to evaluate the performance of their framework.

To minimize the domain gap between synthetic and real images in vehicle re-identification, the authors in [41] developed a re-identification dataset named VehicleX. Objective of the work was to minimize the distribution differences between synthetic and real data using the Fr'echet Inception Distance (FID) metric [42]. To achieve this, the authors propose an attribute descent approach where attributes are optimized iteratively to make the synthetic data more similar to real data. Unity, a 3D engine is used to simulate various vehicle attributes, such as orientation, lighting, and camera parameters. It employs a Gaussian Mixture Model (GMM) to capture attribute distributions and optimize mean values to reduce distribution differences.

Authors in [43] developed a Viewpoint-aware Channel-wise Attention Mechanism (VCAM) that leverages a channel-wise attention mechanism to extract viewpoint-aware features for vehicle re-identification (re-ID) tasks. VCAM focuses on reweighing the importance of each feature map channel based on the viewpoint of the vehicle, aiming to emphasize features from visible parts that are relevant for re-ID matching. It consists of a viewpoint estimation module, VCAM, and a re-ID feature extraction module. The viewpoint estimation module estimates the viewpoint of the vehicle image, generating a viewpoint feature used by VCAM to produce channel-wise attentive weights. These weights refine the feature maps extracted from the re-ID feature extraction module, resulting in a representative feature for re-ID matching. Authors evaluated the framework with VeRi-776 [40] and presented it at the AI City Challenge 2020 [44].

A transformer based vehicle re-identification framework namely TransReID by the authors in [45]. The work highlights the limitations of CNN-based methods in capturing global context and fine-grained details, prompting the need for a new approach. The TransReID framework involves a strong baseline with a pure transformer, a jigsaw patch module (JPM) for rearranging patch embeddings, and side information embeddings (SIE) to alleviate bias due to camera/viewpoint variations. TransReID introduces two novel modules (JPM and SIE) to enhance feature learning and mitigate bias in the context of transformers. The framework is evaluated with existing benchmark datasets for vehicle re-identification.

A Multi-View Spatial Attention Embedding network is designed by the authors in [46] to address the challenges of identifying the vehicles under different viewpoints where the same vehicles may have significantly different appearances. The network consists of a multi-view branch network and spatial attention blocks to learn discriminative features for vehicle Re-ID without requiring extra annotations. The multi-view branch structure focuses on feature learning for specific viewpoints, improving robustness and discriminative power. Spatial attention modules filter cluttered backgrounds and discover subtle local differences. The paper presented several experiments that are conducted on VehicleID [8] and VeRi-776 [40] datasets respectively.

To address the challenges of vehicle re-identification such as viewpoint changes, similar appearances and inter-class similarities, the authors in [47] proposed Deep Features, Camera Views, Vehicle Types, and Colors (DF-CVTC) framework. The proposed method expands the backbone of ResNet-50 [29] by incorporating three subnetworks for extracting view, type, and color-specific features, which are then fused to create informative representations for vehicle Re-ID. The study progressively introduces the view, type, and color subnetworks, showing consistent performance improvements compared to the baseline ResNet-50 model [29] By incorporating these attributes, the method aims to reduce inter-class similarity and improve the discriminative capability of the model. Furthermore, the paper also introduces a progressive learning approach during the training to make the underlying network adapt to vehicle appearance changes across different cameras.

A viewpoint adaptation network (VANet) and a cross view distance metric is proposed by the authors in [48] to address the challenge of identifying vehicles from different viewpoints by integrating multi-level information and denoising generated samples. A cross-view label smoothening regularization (CVLSR) technique is used to assign labels to generate cross-view images based on color domains. The cross-view distance metric module is developed to combine original and generated viewpoint information for multi-view matching. VANet is evaluated on two large vehicle Re-ID datasets, VeRi-776 [40] and VehicleID [8], demonstrating superior performance compared to state-of-the-art methods.

Authors in [49] proposed a novel approach for vehicle re-identification called Triplet Center Loss based Part-aware Model (TCPM). The objective of the study was to extract discriminative features from local details of vehicles, which can be challenging due to the similar appearances of vehicles. The idea lies in splitting the output feature map into horizontal and vertical branches to capture spatial details of the vehicles. By dividing the feature map into multiple partitions, both horizontal and vertical regions are processed separately to extract detailed local features. To optimize the feature learning process, the authors introduce the Triplet Center Loss function to ensure intra-class consistency and inter-class separability of features. The TCPM method is evaluated on benchmark datasets, including VeRi-776 [40] and VehicleID [8].

Authors in [50] addressed the challenging task of vehicle re-identification (re-ID) in urban surveillance systems. The primary goal is to identify the same vehicles across different surveillance cameras with extreme viewpoint variations. The authors propose a large margin metric learning method to enhance the effectiveness of vehicle re-ID systems. The proposed method focuses on extracting fine-grained and discriminative features to capture subtle details like cosmetic differences, dents, etc., that can distinguish vehicles. To improve the efficiency of the training process, they introduced an invader-defector sampling scheme to identify hard samples close to classification boundaries and a kernelized

re-ranking method to enhance performance. Authors compared the performance of their developed frameworks with public benchmarks datasets VehicleID [8], Veri-776 [40], and VERI-Wild [51] respectively.

A Generalized Multiple Sparse Information Fusion (GMSI) for vehicle re-identification is developed by the authors in [52]. GMSI employs three different deep networks to extract multiple features from coarse to fine views, treating these features as multi-view representations. The method involves ResNet [29] for global features, a Drop Network for robust features, and a Hierarchical Attention Network (HANet) to combine distinctive features from different layers. These features are then fused using Multi-view Discriminant Analysis (MvDA) to transfer them into a common space for effective feature fusion. Further a comparison is performed among different loss such as triplet loss, cross-entropy loss [8], and center loss to emphasize their importance during training.

A novel approach called keypoint aligned embedding model (KAE-Net) for learning pose-invariant image embeddings is presented in the paper [53]. The method leverages keypoint annotations to align embeddings with keypoints, enabling the network to learn representations that are invariant to pose variations. By reconstructing keypoint heatmaps as an auxiliary task, the model can effectively capture pose information during training. The KAE-Net architecture consists of KAE-Blocks, each associated with a specific keypoint. These blocks perform channel rescaling, selection, embedding learning, and heatmap reconstruction, optimizing for both the main embedding task and the auxiliary keypoint heatmap reconstruction task. The model output is a keypoint-aligned embedding, which aids in capturing pose-invariant features. The framework is evaluated on benchmark datasets like VeRi-776 [40], Cars196 [54] datasets.

The study presented in [54] introduces a novel approach called Partner Learning for Vehicle Re-Identification (Re-ID) to address challenges like intra-class variability and inter-class similarity due to diverse viewpoints, illumination, and similar appearances. A multi-branch architecture is developed to extract discriminative and fine-grained information without increasing inference time or computation costs. Knowledge Transfer is performed to facilitate learning between different branches, enabling the global branch to leverage local information effectively. A Hierarchical Structural Knowledge Transfer (HSKT) framework is developed to mine discriminative and identifiable details in an end-to-end manner. The HSKT approach includes attention-based, relation-based, and logic-based knowledge transfer techniques to enhance the network's performance. The approach is evaluated on benchmark datasets like VeRi-776 [40], VehicleID citepliu2016deep, and VERI-WILD [51].

A counterfactual Attention Learning (CAL) is introduced by the authors in [55] for vehicle re-identification. The motivation of CAL is to address the limitations of existing attention models, that tend to focus on biased or irrelevant features due to weak supervision and lack of explicit guidance. CAL works by comparing the effects of observed attention and counterfactual attention on the final prediction. By maximizing the difference between these two, the network is encouraged to learn more effective visual attentions and reduce the impact of biased training data. This approach helps the model to focus on discriminative regions and avoid suboptimal results caused by biased clues. The approach is evaluated across many re-identification approaches pertaining to both person and vehicle entities.

A Local Channel Drop Network (LCDNet) is designed by the authors in [56] for re-identifying the vehicles. The LCDNet is designed to overcome the problem of apical dominance by releasing the constraint of the most significant features. The network consists of two branches: the local feature learning network and the attentive local feature learning network. The former learns the most important features in each part of the vehicle images, while the latter drops some of these important features to prevent the dominance of specific cues. To optimize the re-identification model, a batch ranking loss is introduced to ensure the network learns meaningful features to distinguish between vehicles. Additionally, a re-ranking method is proposed based on multi-distance metrics to improve retrieval results by considering various similarities. The performance of the framework is evaluated using the benchmark dataset VeRi-776 [40] dataset.

To enhance the performance of vehicle re-identification for CCTV surveillance, the authors [57] designed a zone specific vehicle re-identification approach. The framework considers CCTV cameras to be grouped into different strategic zones and the effectiveness of the re-identification is checked by re-identifying the query samples across these subsets of cameras. The study analyses the factors that may affect the performance of re-identification due to the positioning of cameras that are used in monitoring of vehicles for an efficient vehicle re-identification. The study also examines the effectiveness of parameter selection to mine the vehicle instances that is needed in modelling the vehicle re-identification algorithms.

A Dual-relational Attention Network (DRA-Net) for vehicle re-identification is proposed by the authors in the paper [58]. A core module of DRA-Net called Dual-Relational Attention Module (DRAM) that simultaneously captures the importance of feature points in both spatial and channel dimensions to create a three-dimensional attention module. This framework is designed to enhance discriminative features and suppress irrelevant ones, leading to improved performance in vehicle re-identification. To validate the effectiveness of the DRA-Net, the study conducted experiments on the VeRi-776 [40] and VehicleID [8] datasets.

Emerging as a possible alternate to the problems caused by inadequate labeled data in vehicle re-identification is by adopting synthetic vehicle re-identification datasets and GAN based approaches. Data security and privacy are major concerns while collecting a large-scale real-world traffic

surveillance dataset. To resolve this issue, the authors in [41] introduced a synthetic vehicle re-identification dataset known as VehicleX, which was generated using a 3D graphics engine. The potential to achieve competitive accuracy in re-identification tasks while mitigating privacy concerns, the flexibility to manipulate environmental factors to reduce content domain gaps, and the ability to scale up data efficiently are all benefits of using synthetic datasets.

Furthermore, GAN based re-identification models closely reflect real-world situations by producing varied and extremely realistic synthetic vehicle images that mimic changes in viewpoint, lighting conditions, and occlusions. Due to limited labeled data and privacy issues, recent research works lead to the investigation of data generation through GANs. The data generation techniques can be in the form of data augmentation, synthetic images that are either obtained by 3D models or using GANs. Authors in [59] proposed VehicleGAN, a Pair-flexible Pose Guided Image synthesis approach for re-identification. Their approach synthesizes vehicle images in a unified target pose using GAN eliminating the need for geometric 3D data of vehicles that are impractical to obtain in real time scenarios. The authors developed a GAN-based approach named ColorGAN [60]. By employing GANs, the authors aim to generate high-quality, diverse images of vehicles with altered colors, thereby enriching the training dataset without requiring the collection of new images. This method not only increases the amount of training data but also introduces greater variance, ultimately leading to improved detection performance of the CNNs. To address the difficulty in identifying vehicles solely based on single-view images, a multi-View GAN (MV-GAN) is proposed by the authors in [61]. Their framework synthesizes realistic vehicle images from arbitrary views thereby normalizing viewpoints for re-identification. The framework also allows for the generation of multi-view representations that incorporate complementary features from both the original and generated images. Thus, when there is scarcity of real data MV-GAN can effectively create a more comprehensive training dataset, thus enabling better performance in vehicle ReID tasks.

A Distillation Embedded Absorbable Pruning (DEAP) for object re-identification is proposed by the authors in [62]. DEAP addresses the challenge of transferring knowledge from a heavy teacher network to a light student network by combining knowledge distillation and structured network pruning. DEAP uses a Pruner-Convolution-Pruner (PCP) unit to incorporate NP's sparse regularization on extra pruners, resolving the conflict between KD and NP. An Asymmetric Relation Knowledge Distillation (ARKD) method is proposed to transfer feature representation and asymmetric pairwise similarity knowledge without adaptation modules. DEAP simplifies the student network to a Teacher-Like yet Light (TLL) network via re-parameterization. Authors have evaluated the framework with standard object re-identification datasets that include VeRI-776 [40] for vehicle re-identification.

To address the challenge of improving the accuracy of lightweight student networks in object re-identification, the authors in [63] proposed a novel method called Pairwise Difference Relational Distillation (PDRD). The objective of the study was to ensure consistent ranking results between a lightweight student network and a large teacher network, considering that object re-identification primarily revolves around ranking. The paper also provided the concept of pairwise similarity difference knowledge, which involves optimizing the relationship between pairwise similarities obtained from the teacher and student networks. They theoretically prove that minimizing the difference relationship between these pairwise similarities lead to more consistent ranking results. To achieve this, a non-linear pairwise difference relational knowledge loss function is designed to enhance the transfer of knowledge. Authors extensively experimented on four different object re-identification datasets to validate the effectiveness of their proposed framework.

A Multi-axis Interactive Multidimensional Attention Network (MIMA-Net) was developed by the authors in [64] to address the challenges of re-identifying vehicles. The main intuition was to capture fine-grained discriminative information crucial for distinguishing between similar vehicles. MIMA-Net incorporates two core modules: Window-Channel Attention Module (W-CAM) and Channel Group-Spatial Attention Module (CG-SAM). W-CAM focuses on capturing channel attention through spatial interactions, while CG-SAM emphasizes spatial attention by interacting across channels. These modules work together to learn discriminative semantic features in vehicle parts efficiently. Authors extensively evaluated their frameworks with existing vehicle re-identification datasets.

Authors in designed a Text Region Attention Network (TANet) [65] for re-identifying the vehicles. They designed TANet to mitigate the challenges faced in re-identifying the vehicles of the same model due to the under utilization of highly discriminative text regions. TANet integrates global and local information with a specific focus on text regions to improve feature learning. The network captures stable and distinctive features across various vehicle views, focusing on text regions to extract resilient and discriminative features. TANet consists of three main modules: a global feature enhancement model, a text area attention module, and an adaptive multi-feature model. The global feature learning model extracts global appearance features from vehicle images using ResNet50 [29] as the backbone. The text area attention module detects text areas on vehicles to generate local text-area features, enhancing the obtained features. The global feature enhancement model combines global and locally enhanced features to improve the expressiveness of global features and obtain a more discriminative vehicle representation. Authors evaluated their framework with standard available datasets namely VeRI-776 [40], VehicleID [8] and VERI-Wild [51] datasets.

A Pose Apprise Transformer Network for vehicle re-identification (PATReID) is developed by the authors in [66]. PATReID contains a two-stream multi-task neural network that incorporates viewpoint invariant features to compute global features for vehicle ReID. The framework includes side information like camera ID and viewpoint ID to enhance robustness against viewpoint variations. The re-identification framewok comprises of three stages namely: Vision Transformer (ViT) network [31] to extract visual features, a Two-Stream Network (ViT [31]+ ResNet50 [29]) for feature combination, and PATReId that incorporates pose information for enhanced global feature representation used for vehicle ReID and attribute classification. The framework optimizes weights through a total loss that combines ID, color, and type losses.

Authors in [11] developed a re-identification framework to re-identify the vehicles observed by UAVs. For their developed VRAI dataset, a multi-task model with attribute classification and a discriminative parts detection branch is designed to perform re-identification. Further a weighted feature aggregation is employed, showing significant performance improvement over average feature methods. The authors extended their work [67] by developing a novel Orientation Adaptive and Salience Attentive (OASA) model for vehicle re-identification in aerial images and videos. It consists of two key modules: the Orientation Adaptive Dynamic Convolution module and the Transformer-Based Salience Attentive module. The Orientation Adaptive Dynamic Convolution module is designed to extract orientation-invariant features by customizing convolutional kernels for each vehicle instance to handle pattern deformations due to UAV views. On the other hand, the Transformer-Based Salience Attentive module focuses on capturing discriminative vehicle clues by integrating valuable information from discriminative parts annotations through a transformer layer and adaptive attention mechanism. Authors evaluated their framework with existing datasets such as VeRi-776, VehicleID and other datasets.

To address the challenges in viewpoint and scale variations for re-identifying the vehicles observed by UAVs, the authors in [12] developed a dataset namely UAV-VeID. The dataset includes multiple images of each vehicle, presenting challenges in viewpoint and scale variations. To address these challenges, the paper proposes a viewpoint adversarial training strategy and a multi-scale consensus loss to enhance feature robustness for vehicle re-identification. The methodology involves training a feature generator to produce viewpoint-invariant features through adversarial learning and utilizing multiple branches with different dilation rates to extract multi-scale features. The multi-scale consensus loss encourages features from different branches to be similar, improving scale invariance. Authors have also discussed the challenges posed by UAV data collection, including similar backgrounds and illuminations, and suggest strategies like unsupervised training algorithms for data annotation.

To address the challenges such as significant size differences and uncertain rotation variations, a Rotation Invariant Transformer (RotTrans) is developed by the authors in [68]. RotTrans introduces a feature-level rotation strategy to simulate rotation transformations on image patches, enhancing the model's robustness against large rotation differences. The authors evaluated the performance of the framework using VRAI [11] re-identification dataset.

Authors in [69] introduced a re-identification framework based on graph matching to address challenges in identifying vehicles taken by UAV from different viewpoints. The framework comprises three sub-modules: Feature Extraction Module (FEM), Graph Convolution Module (GCM), and Graph Matching Module (GMM). FMM Extracts both global and local features of vehicles. It predicts key point heat maps and feature maps from vehicle images, extracting local features from key points and a global feature to create a hybrid vehicle representation. GCM integrates the topological structure information between key points of vehicles into local features through graph convolution. It merges structural information to enhance feature representation and alignment between multiple views. GMM Aligns key features between graphs by learning a robust alignment metric. Authors evaluated the framework with VRAI [11] and VeRI-776 [40] dataset to demonstrate the effectiveness of the framework.

A self-aligned spatial feature extraction network is designed by the authors in [70] for vehicle re-identification using UAVs. The study addressed the challenges faced in vehicle re-identification for UAV surveillance systems, where vehicles with similar color and type can have variable orientations, making it difficult to extract distinguishing characteristics. To overcome these challenges, the authors propose a self-aligned spatial feature extraction network (SANet). The SANet consists of three branches: a global branch for extracting global features and upper/lower and left/right branches for spatial features in different directions. A self-alignment module is introduced to align input images without annotations, allowing for consistent feature extraction. The network utilizes triplet loss [71] functions for global and spatial features to enforce feature similarity constraints. The authors extensively evaluated their framework with UAV-VeID dataset for analysis.

B. CROSS DOMAIN VEHICLE RE-IDENTIFICATION

Most current vehicle re-identification frameworks utilize supervised learning, leveraging sufficiently well-annotated datasets to train their models. These supervised approaches necessitate complete annotations, a requirement that is often unfeasible due to the abundance of unlabeled data. A significant challenge arises when training and testing data are sourced from different domains, as the re-identification models may not generalize well across these domains. Annotating target images from an unfamiliar domain is both labor-intensive and cost-prohibitive. Consequently, many existing works address this issue through domain adaptation techniques. Domain adaptation involves training a model

for the target domain by utilizing a fully annotated source dataset alongside an unlabeled target dataset. This problem is typically tackled through methods such as cross domain unsupervised transfer learning and progressive learning, which aim to bridge the domain gap and improve model generalization across different data distributions. Numerous studies on cross-domain person re-identification [72], [73], [74], [75], [76], [77], [78], [79] have emerged in recent years. Some of these frameworks have also been utilized to evaluate cross-domain vehicle re-identification algorithms. In addition to these works, there are significant contributions specifically developed for vehicle re-identification, which are discussed below.

Authors in [80] proposed a cross domain adaptation framework namely Dual branch adversarial network (DAN) an image-to-image translation network for vehicle re-identification. The primary purpose of the DAN is to translate images of the source domain to target domain and to avoid the need of extra information in the form of spatio-temporal labels, attribute labels. DAN comprises an encoder and generator branch to extract the features of the vehicles from source domain and to encode the information to generate a new image. DAN uses adversarial training mechanisms to distinguish real target domain images from those images generated by discriminators. This end-to-end framework is beneficial when there is abundant labeled data to translate them into an unknown target domain.

To address the diverse environmental changes across the different datasets, the authors in [81] proposed an Attribute Invariant Visual Representation Network (AIVR-Net) designed to obtain attribute invariant features and facilitate discriminative visual representation learning to perform re-identification. The approach undertaken is inspired by compositional zero-shot learning that considers re-identification as an out-of-distribution generation problem. AIVR-Net comprises of two branches to obtain the global features that are combined with vehicle attribute features. The second branch of AIVR-Net removes the invariant attribute features of the vehicles. Authors evaluated the cross domain re-identification framework with VehicleID [8] and Veri-776 dataset [40] for different cross domain techniques.

An unsupervised re-identification method namely semantic transfer-based collaborative matrix representation (STCMR) is developed by the authors in [82]. The primary purpose of the re-identification framework was to address cross domain vehicle re-identification where the labeled semantic information from the source domain may be incomplete or noisy. This method leverages the idea that a vehicle can be represented as a combination of basis vectors from a dictionary that learns semantic features relevant for re-identification. The dictionary is built using data from both source and target domains, allowing the model to transfer knowledge between them. Both source and target domain data are used to train the dictionary. This allows the model to focus on learning semantic features essential for re-identification, mitigating the impact of domain-specific variations.

A domain adaptation framework containing an image-to-image translation network named vehicle transfer generative adversarial network (VTGAN) and an attention-based feature learning network (ATTNet) is proposed by the authors in [83]. The VTGAN translates images from the source domain to the target domain while preserving the identity information of the vehicle and the ATTNet further improves the domain adaptation by focusing on relevant parts of the images during the training process. The VTGAN contains a content encoder, a style encoder and a decoder. To preserve the identity information from the source domain an attention module is developed in the content encoder. According to [83] the VTGAN is designed for vehicle re-identification that does not require any labels for training the re-identification network. Authors evaluated the framework with existing techniques such as CycleGAN [84] and SPGAN [77] on VehicleID [8] and VeRi-776 [40] datasets.

A progressive learning approach with multi-scale fusion network is proposed by the authors in [85] for vehicle re-identification to address challenges in vehicle re-identification (re-ID) specifically in cross domain scenarios. PLM incorporates a progressive learning approach that utilizes unlabeled target data along with labeled data from a source domain. PLM iteratively refines the model by creating "pseudo target samples" through domain adaptation and incorporating these along with unlabeled data for training. With multi-scale attention network features from different layers of the model are extracted that captures both low-level texture and high-level semantic information from vehicle images. A weighted label smoothing loss assigns weights to pseudo labels based on their confidence to improve the training process. The performance of the PLM re-identification framework is evaluated on the existing benchmark datasets namely VehicleID [8] and VeRi-776 [40] datasets.

An unsupervised domain adaptation (UDA) for vehicle re-identification is expected by the authors in [86] for vehicle re-identification. The framework bridges the gap by extending UDA theories to consider the specific characteristics of re-ID. A self-training scheme is developed that is specifically designed for unsupervised vehicle re-identification. This training scheme leverages the unlabeled target data to enhance the model's performance. The network is initially trained with labeled source domain vehicle images. A target data prediction is performed on unlabelled data to determine the initial guess which is further refined with pseudo labels. The model is then fine-tuned using both the original source domain data and the target domain data with these pseudo-labels. This process is iteratively performed until the model gradually improves its ability to handle the target domain even though it never had labeled examples from that domain.

Authors in [87] designed a re-identification framework namely Vehicle Re-identification using PROgressive Unsupervised Deep architecture (VR-PROUD) to address the

task of re-identification when the amount of labelled data is limited. VR-PROUD tackles vehicle re-identification with a unique two-stage process. The first stage uses a CNN to extract vehicle features. Then, an unsupervised clustering step groups similar vehicles based on these features. These clusters are progressively refined and used to train additional CNNs. VR-PROUD incorporates color information to ensure only high-quality clusters are used for training, leading to faster convergence and improved accuracy. This cascaded approach with unsupervised learning is a novel contribution to the field of vehicle re-identification.

A Progressive adaptation learning (PAL) technique for vehicle re-identification is developed by the authors in [88]. The purpose of the study was to address the limited annotated data and domain bias for a model to be adapted to a target dataset. With Weighted label smoothening (WLS) technique the re-identification framework iteratively updates the model to adapt to the unlabeled domain. A domain adaptation module based on GAN concepts aims to bridge the domain gap between a source and a target dataset by generating synthetic pseudo target samples. To ensure the pseudo labels generated to be near perfect, weighted label smoothening loss is used to estimate the similarities for different pseudo cluster images.

C. MULTI-MODAL VEHICLE RE-IDENTIFICATION

The re-identification frameworks make use of the data that are acquired by either CCTVs or UAVs. These data are mainly of RGB single modality image format. It is challenging to re-identify vehicles in environments where there is lack of ambient light, low visibility at night hours, foggy weather or dark scenes. To address these issues in recent years, researchers have contributed datasets and frameworks by utilizing different modality of images obtained by multi-spectral acquisition sensors. Though these imaging techniques address the above-mentioned challenges, designing a re-identification framework that processes different forms of images becomes even more challenging. Several re-identifications frameworks are designed to fuse different forms of multi-spectral images to enhance the re-identification task. The contributions of these works are summarized in the paragraphs below.

To tackle the limitations of traditional vehicle e-identification methods that solely depends on RGB images in the event of poor lighting and bad weather, the authors in [89] developed a multi-modal vehicle re-identification framework that incorporates additional spectrum of images like Infrared (IR), Thermal Infrared (TIR) to improve the re-identification accuracy. A Heterogeneity-collaboration Aware Multi-stream Network (HAMNet) multi-modality re-identification framework is developed to process different data streams allowing the network to get exposed to different modality-specific features. The modality specific streams are further merged to consider the inherent differences between the two spectrum to obtain enhanced representation of the fused information. The framework is evaluated with their developed RGBN300 and RGBNT100 dataset.

A Graph-based Progressive Fusion Network (GPFNet) using a graph convolutional network is developed by authors in [90] to adaptively fuse the features of multi-modality vehicle images. This end-to-end framework uses a graph convolutional neural network (GCN) to adaptively fuse the multi-modality features. As a data enhancement technique, random modality substitution (RMS) is used to extract rich and finer mixed-modality vehicle features. An efficient graph structure is designed to fuse the multi-modality features that are developed using a two stage principle. Further the work also introduced a loss function designed for GCN to make the model learn the discriminative and complementary vehicle features. Authors evaluated the framework using RGBN300 and RGBNT100 [89] dataset.

A multi-modal vehicle re-identification based upon ViT [31] namely hybrid vision transformer (H-ViT) is developed by the authors in [91]. H-ViT comprises of two modules: model-specific controller (MC) and a model information embedding (MIE) structure. The MC controls how information from each modality is processed by the transformer. MC essentially creates separate branches within the ViT architecture for each modality (visible, near-infrared). The MIE injects information about the modality into the system at the patch embedding layer. By combining these modules, H-ViT effectively learns the complementary information from different modalities, resulting in more robust vehicle re-identification

Authors in [92] proposed a Generative and Attentive Fusion Network (GAFNet) to perform multi-spectral vehicle re-identification. GAFNet contains two key modules: Generative Modality Transition Module (GMTM) that bridge the gap between original spectral data, improving the consistency within the data used for training. The second module Attentive feature fusion module that combines combine features from original and transitional modalities at the feature level, focusing on informative areas within the feature maps. Their approach aims to create a more unified and informative feature representation for vehicles by bridging the modality gap and focusing on crucial information even under varying lighting conditions.

Authors in [93] developed a multi-spectral vehicle re-identification framework namely cross-directional consistency network (CCNet) to address the cross-modality discrepancy observed by multi-spectral images of vehicles. A cross-directional center loss is designed to minimize the discrepancy between modalities and between samples of the same vehicle across modalities. Furthermore, an Adaptive layer normalization is designed to address distributional discrepancies within a single modality. It adjusts the distribution of features extracted from each modality. The authors assessed their framework using the MSVR310 dataset, which they themselves created for multi-spectral vehicle re-identification.

The work presented by authors in [94] addressed the challenges of multi-spectral vehicle re-identification in the case of missing sensor data. A Dynamic Enhancement Network (DENet) is designed to handle the scenarios of missing multi-spectral information specific to RGB NIR, TIR modalities. DENet is a three-branch feature extractor network to process each multi-spectral image to learn multi-modality image features. In the event of missing spectral data, a feature transformation module aims to reconstruct the missing spectral data from the available data using dimensionality reduction or interpolation techniques. Further using dynamic enhancement technique, the weights of the model are updated with missing spectral data. Though the work is designed to address person re-identification, the authors evaluated the framework with the RGBNT300 [89] dataset.

A Gradual Fusion Transformer (GraFT) is developed by the authors in [95]. The framework designed for performing multi-modal vehicle re-identification uses a learnable fusion token to guide and influence the self-attention mechanism across encoders to process modality-specific features. A new training strategy is also proposed that make use of triplet loss [71] function to optimize the feature embedding space to obtain an accurate vehicle re-identification. Authors evaluated the framework with RGBN300 [89] dataset and compared the performance with other multi-modal re-identification methods.

Authors in [96] developed a unimodal concatenation framework for multi-modal re-identification. To address situation of modality laziness where the features learned from multi-modal vehicle images are lately fused during training. To ensure each model focuses on learning the relevant features from modality-specific images, the UniCat trains separate models for each of the modality specific vehicle images and concatenates the encoded representations.

Authors in [97] proposed a novel framework for multi-modal vehicle re-identification to address domain generalization problems using multi-spectral vehicle images. To address the challenge of lack of available training data in all modalities, the authors proposed a meta-learning approach to create a model that can adapt to unseen data from different modalities during inference. The study uses the RGBNT100 [89] dataset, designed for multi-spectral vehicle re-identification, where models are trained on two modalities and tested on the third. The proposed meta-learning framework involves separating source domains into meta-train and meta-test domains, simulating cross domain generalization.

Authors in [98] introduced a novel approach called MutualFormer for multi-modal data representation, focusing on RGB-Depth salient object detection (RGB-Depth SOD) and RGB-NIR object re-identification. The MutualFormer method integrates features from RGB and Depth modalities using a multi-layer mutual attention mechanism, enhancing the representation learning process. The work also introduces cross-diffusion attention (CDA) that defines cross affinities based on individual multi-modality to address domain gaps. MutualFormer effectively captures the dependencies of both intra and inter modalities.

FIGURE 4. Challenges faced during re-identifying the vehicles observed by cross platform surveillance systems. The identical vehicles (a) and (b) taken from MCU-VReID dataset [13] appears to be transformed/augmented version of one another presenting additional challenges in re-identification.

Several multi-modal vehicle re-identification methods are also evaluated with existing cross domain person re-identification frameworks [99], [100], [101], [102].

D. CROSS PLATFORM VEHICLE RE-IDENTIFICATION

Cross platform vehicle re-identification involves the precise identification and matching of vehicles recorded across two different surveillance platforms namely CCTVs and UAVs. This task entails mapping the viewpoints of a vehicle observed by CCTV, which generally captures vehicles from side, front, or rear angles, with UAV imagery that offers aerial or oblique perspectives of vehicles in motion. cross platform vehicle re-identification is significant for enhancing surveillance capabilities through the integration of data from multiple viewpoints taken from two contrasting surveillance platforms, thereby providing a comprehensive understanding of vehicle movements. Cross platform vehicle re-identification presents challenges such as significant appearance variations resulting from differences in viewing angles, resolutions, and lighting conditions across the two platforms. In addition to these, Figure 4 illustrates that the vehicles observed by a platform (CCTV/UAV) appear to be a transformed or augmented form of the other (UAV/CCTV). As this mode of re-identification is recently brought into existence, very few contributions have been made. The summary of these contributions are discussed below.

To address the cross platform vehicle re-identification, the authors designed a vehicle re-identification framework [103] using the concept of homography. Initially a required transformation is estimated that takes the view of a vehicle observed in either of the surveillance platform (CCTV/UAV) to view of another platform (UAV/CCTV). Authors in the study estimated a transformation between the images from CCTV and UAV using homography and further applied this transformation on the images of CCTV to obtain a near approximate look if captured by UAV. Nevertheless, the process of re-identification seems to be difficult because of many considerations. In certain instances, it has been observed that the incorrect estimation of the computed homography, caused by poor projections of comparable

points, hinders the proper transformation of vehicle images for re-identifying vehicles. The authors extended their work by contributing a novel cross platform vehicle re-identification dataset named MCU-VReID for cross platform vehicle re-identification [13]. A cross platform vehicle re-identification network named Multi-scale Feature Fusion Transformer (MSFFT) was developed to learn the features of vehicles at different scales. The MSFFT architecture for re-identification utilizes inception-like layers to capture multi-scale features of vehicles. It also uses transformer multi-head attention layers to learn the semantic relationships between various parts of the vehicles, utilizing the feature maps created by the CNN layers. The vehicle re-identification framework is designed with a two-stage training method. The network is initially exposed to learn the semantic transformation of identical vehicles observed by two contrasting views using self-supervised techniques and further uses this knowledge in the second stage to train the network for re-identification. Table 1 distinguishes the different forms of vehicle re-identification by highlighting their strengths and Challenges. Table 2 lists the notable re-identification works published following year 2019.

E. EVALUATION METRICS

1) VEHICLE RE-IDENTIFICATION METRICS

Vehicle re-identification is equivalent to a problem of retrieving instances, where the aim is to accurately match a query vehicle image with its corresponding images in a gallery. During the process of inference, a re-identification network that has been trained calculates similarity scores between the query image and all the vehicle images in the gallery that are potential matches. The similarity scores are utilized to rank the candidate images, with those possessing higher similarity scores (indicating a stronger resemblance to the query image) being placed at the top of the list. The efficacy of this ranking procedure is assessed using Cumulative Matching Characteristics (CMC), which quantifies the probability that a query instance is accurately recognized within different sizes of candidate sets. For a given rank index 'i' CMC value for rank i is defined as follows:

$$CMC(i) = \sum_{r_k=1}^{i} q(r_k) \tag{1}$$

where r_k denotes the index of the rank. If the query vehicle instance contains multiple ground truth in the gallery set, then mean Average Precision(mAP) is used to measure the overall performance of the re-identification system. For a given query image, the average precision is determined as follows:

$$AP = \frac{\sum_{k=1}^{n} P(k) * gt(k)}{N_{gt}} \tag{2}$$

Here n denotes the number of retrieved vehicle samples, k indicates the vehicles retrieved in accordance with the given rank. $P(k)$ is the precision at cut-off rank k in the recall list and gt(k) is boolean indicator that denotes if k^{th} match is true or false. Further the mean Average Precesion is formulated in Equation (3):

$$mAP = \frac{\sum_{j=1}^{Q} AP(j)}{Q} \tag{3}$$

where Q denotes total number of query images.

2) VEHICLE TRACKING METRICS

Existing object tracking methodologies leverage the Classification of Events, Activities and Relationships (CLEAR) [124] framework to assess performance. Key performance metrics include Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Identity Switch (IDSW), and Identification F1 score (IDF1). MOTA evaluates the overall tracking accuracy, incorporating false positives, false negatives, and identity switches. MOTP measures the precision of the tracked object's positions. IDSW counts the number of times a tracked object's identity is incorrectly reassigned. The IDF1 score, which calculates the ratio of correctly identified detections to the average number of ground truth and computed detections, provides a balanced measure of identification accuracy. These metrics are crucial for evaluating tracking algorithms using CCTV and UAV video datasets, ensuring robust and reliable performance in diverse tracking scenarios. IDF1 score is formulated as follows:

$$\begin{aligned} IDP &= \frac{IDTP}{IDTP + IDFP} \\ IDR &= \frac{IDTP}{IDTP + IDFN} \\ IDF1 &= \frac{2IDTP}{2IDTP + IDFP + IDFN} \end{aligned} \tag{4}$$

where IDP, IDR are the identification precision and recall metrics that make use of true positive ID (IDTP) true negative ID (IDTN) and false negative ID (IDFN).

III. VEHICLE TRACKING

Vehicle tracking is essential in the context of autonomous driving [125], [126] and ITS as it serves as the foundation for real-time navigation, traffic management, and safety assurance. Vehicle tracking is the process of utilizing GPS, LIDAR [127], RADAR, computer vision, and communication technologies to constantly observe and transmit the location and motion of vehicles. Precise vehicle tracking in autonomous driving facilitates the autonomous system's ability to maintain situational awareness, which in turn enables accurate path planning, collision avoidance, and adherence to traffic laws. Moreover, the vehicle tracking data contributes to ITS thereby improving the optimization of traffic flow, minimizing congestion, and enhancing emergency response times. Recent years have witnessed a increased growth towards developing the tracking frameworks using vision, deep learning and sensor fusion techniques that takes and process vast quantities of data produced to forecast vehicle

TABLE 1. Summary of different forms of vehicle re-identification. The table lists the strengths and challenges of single platform, cross-domain, multi-modal and cross platform vehicle re-identification.

V-ReiD Form	Category	Strengths	Challenges
Single Platform	CCTV	Strategic decision can be made in positioning the cameras for effective re-identification.	Limited field of view. Poor image quality in dynamic weather condition. Expensive to scale. Availability of vehicles pose from all the perspectives is not guaranteed in case of occlusion or technical fault.
	UAV	High resolution images/video clips enrich the re-identification models to learn more semantic features of vehicles. Able to gather more information of vehicles perspectives due to its mobility.	Variation in flight altitudes makes it challenging to re-identify vehicles observed at different scales. Ego motion blur.
Cross Domain	Unsupervised Domain Adaption/Progressive Learning	Eliminates laborious data annotation for target dataset Reduce domain bias. Mitigates the impact of label noise in training data.	Adverse domain shift is still challenging. Inefficient when the source and target dataset are imbalanced.
Multi-modal	Different Fusion Techniques	Improved performance under low light conditions or night hours by utilizing NIR & TIR modalities. Complementary form of input with NIR, TIR and RGB enhance re-identification of vehicles.	Challenging at generalizing the multi-modal appearance of vehicles. Computationally expensive.
Cross Platform	Homography based Vehicle re-identification	Re-identification of vehicles is performed just by transforming the vehicles observed by one surveillance platform to another.	Requires initial training of a Homography CNN model to estimate the transformation between cross platform vehicle images. May result in an incorrect transformation when there is a minimal corresponding matching points of vehicles observed by two surveillance platforms.
	Two-stage training approach	Semantic transformation between vehicles can initially be learned using self-supervised approaches without labeling. Able to re-identify or retrieve vehicles across platforms, even if the same vehicle is occluded or absent in one platform. Vehicles are re-identified even if there is a minimum overlap of parts observed by either platform.	Annotating identical vehicles observed across network of CCTV cameras and UAV is challenging. The initial two-stage training is exhaustive.

behaviors and make well-informed choices. In summary, the incorporation of advanced vehicle tracking technologies is essential for ensuring the dependability, security, and effectiveness of autonomous driving and ITS.

Typically, vehicles on the road are in motion, which leads to the effects of ego and relative motion. The vehicles encountered on the road exhibit variability in terms of their size, shape, and color [128]. The on-road environment presents a myriad of challenges, including variations in illumination, background, and scene complexity. Complex shadowing, man-made structures, and pervasive visual clutter can lead to erroneous detections. Additionally, vehicles appear in various orientations, such as preceding, oncoming, and cross traffic. The frequent and extensive scene clutter often obscures full visibility, resulting in partially occluded vehicles. Thus, it is required to accurately localize the vehicles in the scenes to provide timely and advanced notice of critical situations to both human and autonomous vehicles. Vehicle tracking is performed using various techniques such as monocular vision, stereovision, 3D vehicle detection and tracking [129], fusion of monocular vision and stereovision techniques, etc. The research findings using these techniques are available in the studies [130], [131], [132].

In recent years the research towards developing an automated system for Multi-target multi-camera (MTMC) tracking has seen an increase in demand. The application MTMC tracking ranges from city-scale traffic management, crowd analysis, and transportation analysis for intelligent city planning. In the context of traffic management, MTMC poses several challenges such as vehicle variabilities, occlusions,

TABLE 2. Summary of prominent vehicle re-identification frameworks. The table presents vehicle re-identification studies published after 2019. These works are categorized based on single platform, cross domain, multi-modal and cross platform vehicle re-identification.

Year	Reference	Category								
		Single Platform		Cross Domain		Multi-modal			Cross Platform	
		CCTV	UAV	Unsupervised Domain Adaptation	Progressive Learning	Early Fusion	Late Fusion	Middle Fusion	Homography based	Two stage Training
2019	Tang et al. [104]	✓								
2019	Lou et al. [104]	✓								
2019	Lou et al. [51]	✓								
2019	He et al. [105]	✓								
2019	Tang et al. [106]	✓								
2019	Chu et al. [107]	✓								
2019	Khorramshahi et al. [108]	✓								
2019	Wang et al. [11]		✓							
2019	Bashir. [87]				✓					
2020	Liu et al. [109]	✓								
2020	Jin et al. [110]	✓								
2020	Meng et al. [111]	✓								
2020	Chen et al. [112]	✓								
2020	Khorramshahi et al. [113]	✓								
2020	Liu et al. [114]	✓								
2020	Zheng et al. [115]	✓								
2020	Li et al. [89]					✓				
2020	Song et al. [86]			✓						
2021	Sun et al. [116]	✓								
2021	He et al. [45]	✓								
2021	Li et al. [113]	✓								
2021	Rao et al. [55]	✓								
2021	Zhao et al. [117]	✓								
2021	Zhang et al. [118]	✓		✓						
2021	Teng et al. [12]	✓	✓							
2021	Holla et al. [103]								✓	
2022	Zhu et al. [119]	✓								
2023	Wu et al. [120]	✓		✓						
2023	Zhou et al. [121]	✓								
2023	Yao et al. [122]	✓		✓						
2023	He et al. [90]						✓			
2023	Pan et al. [123]							✓		
2023	Zheng et al. [93]						✓			
2024	Yu et al. [34]	✓		✓						
2024	Li et al. [59]	✓								
2024	Holla et al. [13]									✓

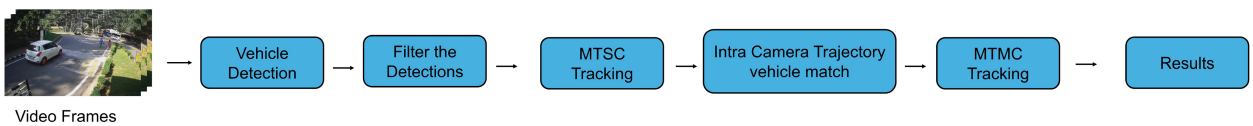

FIGURE 5. The vehicle tracking pipeline starts with vehicle detection, followed by filtering to remove incorrect detections. Within each camera, vehicle tracklets are created using single-camera tracking. These tracklets are then matched with tracklets from neighboring cameras for multi-camera, multi-target tracking.

difference in the appearance of vehicles, etc. Several public datasets are made available for the researchers to provide feasible solutions to enhance the traffic flow required for ITS. Commonly major MTMC tracking frameworks follow the tracking-by-detection pipeline provided in Figure 5. Given the video feeds from the surveillance cameras, the vehicle detection is performed on each frame using handcrafted feature extraction computer vision techniques or with advanced techniques such as CNNs or transformers. These detections are filtered out to eliminate the spurious detections to reduce the occurrence of false positives. With valid detections the intra-camera vehicle trajectory matching aligns and merges vehicle trajectories that may have been fragmented due to occlusions or detection failures, ensuring a coherent path for each vehicle within each camera view. The vehicle tracklets obtained at individual surveillance cameras are associated across neighboring cameras to perform multi-target multi camera tracking. Potential techniques such as appearance features, spatial-temporal constraints, and potentially pre-calibrated camera geometry are used to maintain consistent tracking vehicle identities as they move between cameras. Though several contributions are provided towards object detection [133], [134], [135], [136], [137], the present study limits the discussion of vehicle tracking presented using CityFlow [104] dataset, AI City challenge workshops [44], [104], [138], [139], and other contributions that involves UAVs.

A multi-camera vehicle tracking system called ELECTRICITY for intelligent city and traffic management is designed by the authors in [140]. The system aims to track vehicles across multiple surveillance cameras, overcoming challenges like appearance variance due to viewing perspectives and synchronization of tracking IDs across cameras. The proposed system consists of several modules, including object detection, multi-object tracking, multi-camera re-identification, and tracking ID

synchronization. Detection techniques such as Mask-RCNN [141] and an online multi-target tracking algorithms like Deep SORT [142], and object re-identification using aggregation loss with triplet loss and hard mining are adapted for the underlying task. The performance of the framework is evaluated on the AI City 2020 [44] Track 3 dataset, which includes various scenes and cameras for training and validation.

The authors in [143] focused on developing automated systems for intelligent transportation in smart cities. The paper addresses the challenges of vehicle re-identification, multi-camera vehicle tracking, and anomaly detection. As a part of their work, an Excited Vehicle Re-identification (EVER) framework is developed that uses a self-supervised residual generation technique to enhance representation learning. An excitation layer is incorporated during training to excite intermediate feature maps and improve discriminative vehicle representations. To facilitate multi-camera tracking, a single camera tracking is applied using DeepSORT [142] and MOANA [144]. Their tracking framework employs a distance matrix computation and clustering mechanism to obtain multi-camera vehicle trajectories efficiently. The re-identification and tracking framework are evaluated with the tracks provided in AI City Challenge dataset [44].

To address vehicle counting task at junction points, the authors in [145] addressed the problem as vehicle tracking task that involved multiple vehicle tracking using Yolo [146] and DeepSORT [142] for detection and tracking, respectively. Distinguished regions were defined based on geographical features and movements in each scenario. By linking regions and considering conflict regions, they aimed to enhance vehicle monitoring and counting accuracy. Authors evaluated the framework with AI City Challenge dataset [44].

An MTMCT task is addressed by the authors in [145] to tackle the challenges such as generating local tracklets for all targets under each camera and matching these local tracklets across different cameras to form complete global trajectories for each target. Their approach consisted of three key phases, i.e., local tracklet generation, semantic attribute parsing and tracklet-to-target assignment. Authors evaluated the framework with AI City challenge dataset [44].

Authors in [147] address the challenging task of vehicle turn-counts by class at multiple intersections using a single floating camera. An integrated algorithm is proposed that combines detection, background modeling, tracking, trajectory modeling, and matching in a sequential feedback cycle. The approach involves using a GMM like background modeling technique for detecting moving objects, which is then combined with a deep learning-based detector for high-quality vehicle detection. A multi-object tracking method is employed to generate trajectories for each vehicle. Trajectory modeling and matching are used to leverage the direction and speed of local vehicle trajectories for accurate turn-counts at intersections.

Using the crossroad zone information authors in [148] developed a multi-camera vehicle tracking framework. The framework makes use of vehicle detection, Re-ID feature extraction, single-camera tracking, and multi-camera trajectory generation. Their framework comprises of several modules such as Tracklet Filter Strategy (TFS), Direction Based Temporal Mask (DBTM), and Sub-clustering in Adjacent Cameras (SCAC) are introduced to address challenges in vehicle tracking. TFS is used to filter tracklets and improve precision without external object detectors. The Direction Based Temporal Mask (DBTM) reduces the matching space for visual re-identification by considering the moving directions of vehicles in different zones. sub-clustering in Adjacent Cameras (SCAC) matches tracklets in adjacent cameras to handle appearance changes in vehicles. These modules collectively enhance the tracking performance. Authors evaluated the performance of the framework with AI City Challenge dataset [138].

A multi-camera vehicle tracking system is developed by the authors in [149] for city-scale traffic management. The system consists of three main components: single-camera tracking (SCT), vehicle re-identification, and multi-camera tracklets matching (MCTM). Their framework aims to overcome challenges such as occlusion, similar vehicle models, and feature variations due to lighting and perspective differences. In the SCT phase, the system employs a tracker update strategy and a template matching method to enhance tracking precision. The vehicle detector EfficientDet [150] is used for effectiveness. For re-identification, a model is employed to robustly extract appearance features, and a spatial attention mechanism is implemented. Synthetic data is utilized to train attribute re-identification models, improving features extraction. The paper addresses imbalanced re-identification data by employing data augmentation methods and spatial attention. It proposes an efficient tracklets representation strategy and uncertainty-based matching for MCTM, utilizing spatial-temporal information to enhance tracking accuracy. Authors evaluated their frameworks with VeRI776 [40] and AI City Challenge [138] datasets.

A Candidates Intersection Ratio (CIR) to evaluate the generation of vehicle tracklets in MTMC is proposed by the authors in the study [151]. Their approach consisted of four modules namely vehicle detection using Faster-RCNN [152], vehicle detection association, single camera tracking and multi-camera vehicle tracklet matching. Different other techniques such as IBN-Net, a method that combines instance normalization (IN) and batch normalization (BN) to enhance image appearance modeling in the network. For tracking, strategies such as filtering strategies in SCT to rule out unsuitable vehicle tracklets, including speed filtering, stay time filtering, and IoU filtering. The CIR tracklet association algorithm iteratively considers all possible pairwise matching across camera views. It constructs a binary tree-like structure to uniquely identify vehicle IDs and avoid cyclic associations, ensuring robust tracklet association. Authors evaluated the framework with AI City challenge [138] dataset.

To address the primary challenges faced in multi-camera vehicle tracking, authors in [153] proposed a solution where a model is trained with multi-loss to extract appearance features and a filter with spatial-temporal information between cameras. Their approach involves three main steps: generating tracklets in a single camera through vehicle detection and multi-target tracking, extracting appearance features through a trained vehicle ReID model, and proposing a matching strategy that considers appearance feature similarity, time information, and space information between adjacent cameras. Techniques such as SORT and DeepSORT [142] algorithms, combined with feature similarity matching are used to enhance the tracking performance. Authors evaluated the performance of the framework with AI City Challenge dataset [138].

A MTMC vehicle tracking framework is proposed by the authors in their work [154]. Their framework is divided into three main stages: vehicle detection and feature extraction, multi-target single-camera tracking, and multi-camera association of local trajectories. For vehicle detection, the Mask R-CNN [141] method is used, followed by pre-processing steps to refine the detections. Appearance features are extracted for each detected object using a ResNeXt-50 [155] network, which is trained using SGD with Nesterov momentum and triplet loss. The feature extractor is initialized with ImageNet pre-trained weights and augmented with data for robustness. The multi-target single-camera tracking involves connecting trajectories between frames by comparing appearance features and using the Hungarian algorithm [156] for association. Post-processing steps are applied to remove stationary trajectories and connect moving trajectories. In the multi-target multi-camera tracking phase, trajectories from each camera are matched using pairwise Euclidean distances and the Hungarian algorithm. Trajectories are connected across cameras based on appearance similarity and spatial-temporal information. The performance of the MTMC framework is evaluated with AI City Challenge dataset [138].

A multi-target multi camera tracking framework is designed by the authors in [157]. The framework involves an occlusion handling strategy for single camera tracking to remove false detections and verifying assignments to non-moving vehicles. These features enhance performance under occlusion, especially in cluttered regions and bypassing scenarios prone to tracking errors. For multi camera tracking, the framework incorporates a hierarchical clustering method based on vehicle re-identification features, using a scene model, topological and temporal constraints, and a background filtering component. Vehicle re-identification module employs a global feature learning approach to re-identify the vehicles across different scenes.

A vehicle tracking system in city-scale camera network is designed by the authors in [158]. The tracking system focuses on single-camera tracking, appearance feature re-identification, and multi-target multi-camera tracking. For SCT, Mask R-CNN [141] is used for object detection, and TrackletNet Tracker (TNT) [159] generates trajectories based on temporal and appearance info. The proposed Tracklet Reconnection technique refines tracklets affected by occlusion using zone areas and GPS data. In re-identification, the paper addresses the data imbalance issue when using large auxiliary datasets to train on smaller main datasets. The Balanced cross domain Learning (BCDL) approach is introduced to train a model on both datasets, avoiding overemphasis on the auxiliary set. The training batch consists of a fixed rate of data from each dataset. The system uses a feature extractor trained with BCDL to improve re-identification performance. The Tracklet Reconnection technique improves the initial SCT results by associating split tracklets. In the proposed method, the completion scores and inherent constraints are used to filter invalid tracklet pairs. GPS data enhances candidate pair quality. BCDL mitigates data imbalance between main and auxiliary datasets for re-ID, improving model performance. The designed system for tracking is evaluated with AI City challenge dataset [138].

A robust MTMC tracking system is designed by the authors in [160]. Their system consisted of four key stages: vehicle detection, re-identification, multi-target single-camera tracking, and Inter-Camera Association (ICA). For detecting vehicles, the system uses Cascade R-CNN [161] with a ResNet-101 [29] backbone and Feature Pyramid Network (FPN) for feature extraction. To match the vehicles across neighboring cameras HRNet [162] and Res2Net [163] feature extractors are used. Loss functions like triplet loss and circle loss are employed, and synthetic data generation methods like GANs are explored to enhance Re-ID performance. For multi-target single-camera tracking a Tracklet-Plane Matching (TPM) algorithm to track multiple vehicles within each camera. The ICA stage associates candidate trajectories across successive cameras using a distance matrix refined by constraints like traveling time, road structures, and traffic protocols. The authors evaluated the performance of the tracking system with AI City challenge dataset [138].

To address the challenges of low-confidence object detection and precise trajectory association, the authors in [164] developed a multi-camera multi-vehicle tracking framework. The tracking technique used in the paper involves a cascaded tracking method that leverages detection, appearance features, and trajectory interpolation to improve detection and identification recall. In the cross-camera tracking phase, a zone-gate and time-decay based matching mechanism is introduced to optimize the matching space, ensuring high identification precision. Space, time, and appearance features are crucial for trajectory association, and the proposed methods adjust the appearance matrix to link tracklets accurately across different cameras. A trajectory post-processing technique is also presented to enhance the coherence of tracking results. Authors evaluated the framework with CityFlow [104] and AI City Challenge [139] datasets.

To address the challenges observed in multi-target multi-camera tracking due to occlusions and identity switches,

the authors in [165] developed a framework to mitigate these deficits. The framework consists of several modules, including tracklet splitting, clustering, and completion, to enhance single-camera tracklets and facilitate cross-camera association. The Track Refinement Module (TRM) refines single-camera tracklets by filtering false detections, splitting tracklets based on direction changes, and using K-Means clustering for identity switches caused by multiple vehicles. The TRM also includes a track completion component to connect track fragments and correct tracking errors, ensuring that tracklets are accurately merged. The framework has been evaluated on CityFlowV2 [104] dataset and AI City challenge dataset [139].

A multi-camera vehicle tracking system was developed by the authors in [166]. Their system comprises of four components that include detection and re-identification models, single-camera tracking enhancements, zone-based single-camera tracklet merging, and multi-camera spatial-temporal matching and clustering strategy. For vehicle detection, the authors utilize the YOLOv5 model pre-trained on the COCO dataset, focusing on detecting cars, trucks, and buses. They also employ ResNet [29] models for vehicle re-identification. In single-camera tracking, the authors enhance the SORT [167] framework by introducing augmented tracks prediction, multi-level association, and a feature dropout filter to address broken tracklets and ID switches. They also propose a zone-based tracklet merging strategy to link tracklet fragments within a single camera. For multi-camera tracking, the authors develop a spatial-temporal matching strategy that reduces the search space during matching and improves hierarchical clustering to capture complex tracklet scenarios like U-turns. By clamping trajectories based on GPS location and applying trajectory clamping masks, they significantly enhance the accuracy of ID matching across cameras. Authors evaluated their frameworks using AI City challenge dataset AI City Challenge dataset [139].

A multi-camera multi-target vehicle tracking system is developed by the authors in [168]. They designed the system to address challenges such as lack of labeled data, distortion in vehicle appearances, and ambiguity between similar vehicles. The system consists of three major components: a motion-driven vehicle tracker for robust trajectories, TransReID [45] with MixStyle for domain generalization, and contextual constraints like neighbor matching to resolve appearance ambiguity. Object detection is performed using YOLOv5x. The motion-driven tracking uses Kalman Filter-based predictive estimations for slow and fast vehicles. The vehicle re-identification module utilizes the TransReID model with MixStyle [72] to leverage real and synthetic data and enhance domain generalization. Triplet loss [71] and cross-entropy are used for supervised contrastive learning. The framework is evaluated with the AI City challenge dataset [139].

To address the challenges of tracking such as detection noises, broken tracklets, occlusion, and ID switches, the authors in [166] designed a tracking framework. The framework uses YOLOv5 for localizing the vehicles, ResNet feature extractors to perform re-identification across cameras, SORT [167] and Efficient Convolution Operators (ECO) [169] for single camera vehicle tracking. The multi-camera vehicle tracking involves spatial-temporal matching and aggregation strategies to reduce search space and solve challenges like U-turns, along with a zone-based tracklet merging approach. Authors evaluated the framework with AI City challenge dataset [139].

A multi-camera vehicle tracking framework based on occlusion-aware and inter-vehicle information is designed by the authors in the study [170]. The proposed framework consists of four modules: vehicle detection and feature extraction, single-camera tracking, tracklets similarity analysis with prior traffic knowledge and inter-vehicle information, and clustering for multi-camera tracklets matching. The occlusion-aware module addresses occlusion challenges by segmenting tracklets of occluded vehicles. It recalculates similarity to reduce occlusions and false detections. The inter-vehicle information module enhances tracklet matching accuracy between different cameras, distinguishing similar vehicles. Authors evaluated the framework with AI City Challenge [139] dataset to validate the effectiveness of the framework.

Authors in [171] designed a multi-camera multi-target vehicle tracking system designed to address challenges in tracking vehicles across multiple cameras. Key techniques such as DeepSort [142], Kalman Filter [167] for Single-camera multi-target tracking and K-reciprocal nearest neighbor algorithm for associating candidate trajectories between neighboring cameras. The performance of the framework is evaluated with AI City Challenge dataset [104].

To perform Multi-Target Multi-Camera Tracking (MTMCT), the authors in [172] developed a framework that contains several modules. A Traffic-Aware Single Camera Tracking (TSCT) that detects traffic-aware zones within each camera using MeanShift clustering and performs single-camera ReID to reconnect isolated trajectories caused by traffic scenarios. A Trajectory-Based Camera Link Model (CLM) that establishes relationships between adjacently connected cameras by defining entry/exit zones and transition times. A re-identification module with temporal attention that aggregates frame-level features with clip-level features. A hierarchical clustering technique is finally used to merge vehicle trajectories across all cameras to obtain the final MTMCT results. Authors evaluated the framework with standard benchmark CityFlow dataset [104]

A Multi-Stream Siamese and Faster Region-Based Network (MSRT) is proposed by the authors in [173]. The MSRT combines a multi-stream Siamese network for efficient target search and template updates with a Faster R-CNN detector for object re-identification using object category information. This allows it to address the challenges such as significant target appearance changes, occlusion, and out-of-view scenarios, improving tracking performance.

MSRT involves two main modules: the multi-stream instance search module (MS) and the re-identification and location module (RL). MS utilizes a multi-stream Siamese network to search and update instance templates, generating response maps and target candidates. RL, on the other hand, employs a metric learning model and a Faster R-CNN detector to re-identify and locate lost targets based on the tracking state estimated by MS.

To address the task of multi-target multi-camera tracking, the authors in [174] proposed a framework that consists of two components: single camera tracking (SCT) and inter-camera tracking (ICT). For SCT, a traffic-aware single camera tracking (TSCT) method to address the issue of isolated tracklets. TSCT leverages a TrackletNet tracker [159] and MeanShift clustering to generate traffic-aware zones and reconnect isolated trajectories. The First-In-First-Out (FIFO) strategy is employed for trajectory merging. In the ICT phase, a re-identification model and metadata classifier is used to create appearance and metadata features for each trajectory. A trajectory-based camera link model (TCLM) was established to incorporate spatial and temporal information for improved multi-camera tracking performance. The TCLM handles transitions between cameras, generates zones, and estimates transition times. The framework is evaluated with standard benchmark CityFlow dataset [104].

By using motion information, the authors in [175] proposed a motion-based tracking approach called Local-Global Motion (LGM) tracker for vehicle tracking. The LGM tracker consists of two stages: box embedding and tracklet embedding. In the box embedding stage, boxes are associated into tracklets using deep graph convolutional neural networks (GCN). The tracklet embedding stage further associates tracklets into tracks by learning global motion consistencies. This stage utilizes a reconstruct-to-embed strategy with an attention-based GCN to capture both local and global motion patterns. The framework is evaluated the KITTI [176] and UA-Detrac [177] benchmark datasets.

The authors in [178] discussed a framework for monocular quasi-dense 3D object tracking, focusing on autonomous driving scenarios. The key contributions include integrating quasi-dense similarity learning into the tracking framework, enhancing depth-based data association with centroid-based similarity, and improving the VeloLSTM module to model object velocity, heading angle, and dimension. Various datasets, such as nuScenes [179] and Waymo Open [180], are used for evaluation, demonstrating the robustness and scalability of the proposed pipeline in real-world driving scenarios.

Authors in [181] introduced a new multi-vehicle multi-camera tracking dataset namely Synthehicle to address the challenges of accurately localizing and tracking vehicles in 3D from multiple cameras. Synthehicle is a synthetic dataset that includes ground truth annotations for 3D bounding boxes, depth estimation, and instance, semantic, and panoptic segmentations. Their vehicle also highlights the correlation between single-camera performance and multi-camera tracking performance, emphasizing the impact of scene density and the quality of single-camera tracklets on multi-camera tracking results. The results show how different weather conditions affect tracking performance and the importance of training data diversity for model generalization.

A long-term tracking framework called CRT is developed by the authors in [182]. The CRT consists of a tracker, detector, and classifier, to address challenges in object tracking such as occlusions, scale variations, and target re-detection. The framework integrates a tracker that estimates the target's position, a detector that re-detects the target in case of failure, and a classifier that identifies the correct target. The classifier is a Multilayer Perceptron trained offline on a dataset and predicts the target box based on features extracted from candidate boxes. The performance of the framework is evaluated with benchmark datasets such as UAV123, UAV20L [183] and TLP [184].

A Graph Tracking with re-identification (GTREID) framework for multi-object tracking is developed by the authors in [185]. The GTREID framework combines detection, association, and re-identification features to improve tracking accuracy, particularly under challenging conditions like occlusion in aerial-based vehicle tracking scenarios. The approach integrates data augmentations, appearance features from LABNet [186] and FairMOT, and a graph neural network for robust tracking. The network architecture involves a detector based on FairMOT [187], re-identification features from CenterNet [188], and an association network leveraging Sinkhorn Normalization for soft assignments. GTREID incorporates class-based triplet loss during training to enhance the discriminating capabilities of the centerpoint appearance feature. The dataset used for evaluation is the UAVDT [189] benchmark, known for aerial vehicle tracking under various conditions.

IV. VEHICLE RE-IDENTIFICATION AND TRACKING DATASETS

A. COMPCARS

To perform vehicle categorization and classification, authors in the study [190] presented a large-scale dataset called CompCars. The dataset consists of images from two scenarios: Web-nature and surveillance-nature, providing a diverse and extensive collection of car models. It includes viewpoints (front, rear, side, front-side, and rear-side), car parts (exterior and interior), and rich attributes like maximum speed, displacement, door number, seat capacity, and car type. It contains 214,345 images of 1,687 car models. The CompCars dataset offers unique features like car hierarchy, attributes, viewpoints, and car parts. It organizes car models into a tree structure based on make, model, and year, enabling fine-grained analysis. The dataset includes data from Web-nature sources like forums and public websites, as well as surveillance-nature data captured by cameras, providing a diverse and challenging dataset for cross-modality analysis.

FIGURE 6. Sample vehicle images from VeRi-776 dataset.

B. VEHICLEID

The VehicleID dataset was proposed by the authors in [8] to simplify the process of re-identifying the vehicles. This dataset comprises 221,763 images of 26,267 vehicles that were captured by surveillance cameras in a small city in China. 10,319 vehicles with 90,196 images are labeled with vehicle model information in the VehicleID dataset. There are only 250 vehicle models that have been most frequently seen. Some of these models have over 200 vehicles, while others are rare and have only one vehicle. The dataset is divided into two sections for the purpose of model training and testing. The initial part comprises 110178 images of 13134 vehicles, and 47558 images have been labelled with vehicle model information. The second part includes 111585 images of 13133 vehicles, and 42638 images have been labelled with vehicle model information.

C. VERI-776

A large-scale vehicle re-identification dataset called VeRi-776 is contributed by the authors in [40] for vehicle re-identification. The dataset is developed using 20 surveillance cameras consisting of 51,035 images of 776 different vehicles. These data are acquired from different scenes that include two lane, four lane and cross roads. The images of each vehicle are captured from 2$\tilde{1}$8 viewpoints with variation in illumination and brightness. Furthermore, annotations in the form of license plate and spatial temporal relations are provided for the vehicles appearing in different tracks. This dataset is widely used by different researchers to evaluate the vehicle re-identification frameworks. Figure 6 illustrates some of the sample images of the vehicles taken from VeRi-776 dataset.

D. VEHICLEX

A large-scale synthetic dataset generator called VehicleX for re-identification is developed by the authors in the study [41]. The dataset contains 3D models of various 1,362 vehicles. The dataset generator has a range of backbone models and textures to adapt for different types of vehicles. The dataset generator could also be used for other vision related tasks such as semantic segmentation, object detection, 3D re-construction, etc.

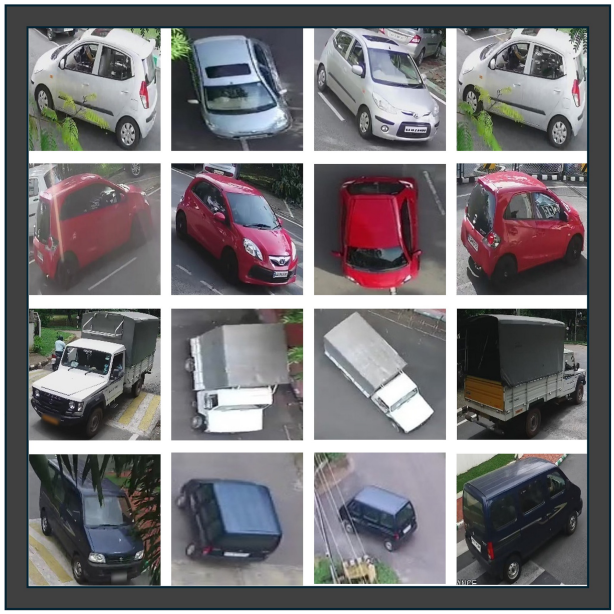

FIGURE 7. Sample images from MCU-VReID dataset.

E. MCU-VREID

A novel dataset for cross platform vehicle re-identification called MCU-VReID is developed by the authors in [13]. The dataset is acquired using 42 CCTV cameras and a UAV to facilitate the cross platform vehicle re-identification of 51 identical vehicles in 5,630 image keyframes [191], [192]. The dataset does not provide supplementary information beyond vehicle and camera identities, making the task of re-identification more challenging. MCU-VReID introduces additional complexities, such as significant transformations in vehicle appearance between CCTV and UAV views, limited overlapping parts in vehicle images, and varying scales of vehicles captured by UAV. Figure 7 illustrates few vehicle samples taken from MCU-VReID dataset.

F. CITYFLOW

CityFlow, a multi-target multi-camera vehicle tracking and re-identification dataset for urban environments is contributed by authors in [104]. The dataset consists of over 3 hours of HD videos from 40 cameras across 10 intersections in a mid-sized U.S. city, making it the largest dataset in terms of spatial coverage and camera numbers for urban traffic analysis. The dataset includes more than 200,000 annotated bounding boxes covering various scenes, viewing angles, vehicle models, and traffic conditions, enabling spatio-temporal analysis. Privacy concerns are addressed by redacting license plates and human faces in the videos. The dataset is divided into five scenarios, with 229,680 bounding boxes of 666 vehicle identities annotated, each passing through at least two cameras. The

FIGURE 8. Sample images form CityFlow dataset.

videos have a resolution of at least 960p and a frame rate of 10 FPS, allowing for synchronization. Annotations were carefully labeled to support trajectory-level annotation across cameras. The dataset provides camera geometry and calibration information for precise spatial localization. The dataset is designed to facilitate research in MTMC vehicle tracking, addressing challenges such as varying viewpoints, fast vehicle speeds causing motion blur, and overlapping field of views between cameras. Figure 8 illustrates vehicle samples provided in CityFlow dataset.

G. RGBN300

A multi-spectral vehicle re-identification dataset is developed by the authors in [89]. The multi-spectral dataset, namely RGN300 contains RGB and NIR vehicle images of 300 identities from eight camera views. The dataset comprises a total of 50,125 RGB images and 50,125 NIR images. Additionally, they acquired TIR (Thermal Infrared) data for 100 vehicles to create a three-spectral dataset for vehicle Re-ID. The dataset presents re-identification challenges such as occlusion, view changes, glaring and abnormal lighting. The authors have additionally captured 17250 thermal infrared images of 100 vehicles from RGBN300 dataset. An additional dataset is designed namely RGBNT100 is designed that contains supplemental TIR data and corresponding RGB-NIR image pairs with 17250 image triples for re-identification. Each of these three camera modalities are acquired with 25 fps. Figure 9 presents the multi-spectral form of vehicle images that are provided in RGBN300 dataset.

H. MSVR310

The authors in the study for multi-spectral vehicle re-identification contributed a MSVR310 re-identification dataset [93]. The dataset contains 2087 samples of 310 identical vehicles. The instances of vehicle images vary from 2 to 20. The dataset is designed in various environments scenarios such as significant illumination changes, occlusion, weather changes, etc. In contrast to the multi-spectral re-identification dataset developed in [89], the MSVR310 provides time labels to match the identical vehicle samples observed in same viewpoint.

FIGURE 9. Sample images from RGBN300 dataset.

I. VRAI

To explore re-identification in aerial videos, the authors in [11] developed a vehicle re-identification dataset named VRAI. The data is collected using two DJI Phantom4 UAVs equipped with 350 sets of video clips, containing 13,022 identical vehicles. Various types of attributes are available for vehicle instances to improve the training task. The cameras of two UAVs capture images of each vehicle instance at different locations, encompassing a range of view-angles and flight-altitudes. In addition, the annotations are provided to identify the distinctive components that aid in differentiating a vehicle from others in each vehicle image. The VRAI datasets present additional difficulties for vehicle re-identification due to the presence of vehicles with a greater range of poses and resolutions. Figure 10 provides vehicle images taken by UAVs in VRAI dataset.

J. UAV-VEID

A dataset called UAV-VeID for vehicle re-identification, obtained using UAV surveillance devices, has been provided in [12]. The dataset was obtained by utilizing two UAVs and consists of 4,601 vehicles, 41,967 annotated images of vehicles, and 16,850 query images specifically focused on vehicles. Each vehicle image is captured from distinct perspectives. Furthermore, the dataset includes a supplementary collection consisting of 300K erroneously identified bounding boxes and vehicles that are not part of the 4,601 annotated vehicle identities.

FIGURE 10. Sample images from VRAI dataset.

K. SYNTHEHICLE

A synthetic dataset called Synthehicle for detecting tracking and segmentation of vehicles is developed in the study by the authors in [181]. Synthehicle consists of 17 hours of labeled video information acquired from 340 cameras. The data is recorded for 64 different days with various scenarios depicting day, night, dawn and rain scenes. They have designed synthetic dataset using an open-source simulation tool called CARLA [193] for building urban traffic scenarios. The research establishes two scenes for eight towns: one for a non-overlapping setup and the other for an overlapping camera view. Vehicles and pedestrians are randomly generated, and cameras are positioned to resemble real-world locations. Information regarding all vehicles within the camera's field of view is acquired by semantic rotating LIDAR sensors. Scenes are recorded in four distinct ambience configurations: day, dawn, rain, and night. The dataset is designed to perform tracking of vehicles across overlapping and non-overlapping cameras with benefits of both 2D and 3D bounding boxes.

L. UAV123

An aerial video dataset for low altitude UAV target tracking called UAV123 is developed by the authors in [183]. The dataset comprises of 123 fully annotated HD video sequences that are acquired at low-altitude settings. The dataset is divided into 3 subsets that contain objects observed by UAV at 5-25 m at different frame rates ranging from 30 to 96 FPS and the resolutions between 720p to 4K. The set 2 of UAV123 dataset contains 12 sequences captured from UAV that are acquired with low resolutions and contains noise due to limited video transmission bandwidth. Set 3 contains 8 synthetic sequences captured using a UAV simulator tool that is rendered using Unreal4 Game Engine. The dataset serves as the primary purpose to perform target detection and tracking.

M. UAVDT

A UAV dataset called UAVDT is developed by the authors in the study [189] for object detection, single object tracking and multiple object tracking. The dataset consists of 100 video sequences that are collected by UAV for 10 hours at urban areas containing scenes from arterial streets, toll stations, highways, crossings and T-junctions. The acquired videos are taken at 30fps at a resolution of 1080 $\times$ 540 pixels. Authors have provided annotations for 2,700 vehicles containing 0.84 million bounding boxes. To perform multi-object tracking, three attributes were considered: weather condition, flight altitude and view of the camera. The data to perform single-object tracking also contains additional set of information such as background clutter, camera rotation and small objects.

Significant contributions have been made over the years to address vehicle re-identification and tracking by developing the datasets. The re-identification datasets are specifically designed for either of the re-identification modes (Single Platform, Cross-domain, Multi-modal and Cross platform) provides adverse challenge for developing re-identification algorithm. Vehicle tracking datasets are developed by providing a short video sequence that are acquired by multiple surveillance or aerial devices exhibit drastic changes in background, motion, scale, etc. Current study provides an exhaustive summary of the prominent vehicle re-identification and tracking datasets that are more commonly used to measure the performance of the vehicle re-identification and tracking frameworks. Table 3 summarizes the prominent vehicle re-identification and tracking datasets.

V. CHALLENGES AND POTENTIAL RESEARCH PATHWAYS

Vehicle re-identification and tracking is a vital process that allows for the effective identification of the same vehicle across multiple cameras. This ability is especially crucial for the purposes of traffic management and security. Extensive research has been dedicated to addressing obstacles such as occlusion, variations in vehicle appearance caused by different viewpoints, camera movement, and differences in scale. Researchers address these challenges by creating varied datasets, advanced feature extractors, novel training strategies, and integrating additional information to achieve precise vehicle re-identification. Additionally, various approaches have been explored by integrating multiple forms of input, which adds complexity to the task of vehicle re-identification and tracking. This study highlights the remaining challenges in this field and suggests potential research directions to address these unresolved issues.

A. DATA AVAILABILITY CONSTRAINTS

In order to create a re-identification and tracking framework, it is essential to have access to extensive datasets that encompass a wide range of challenges for researchers to tackle. Despite notable contributions towards dataset development, these datasets are not accessible to the public due to privacy concerns. Privacy regulations during data acquisition may impose restrictions on the level of detail of vehicle information that can be captured by public cameras.

TABLE 3. Summary of existing vehicle re-identification and tracking datasets. The datasets are categorized based on different modes of re-identification, input format, number of identical vehicles, data acquisition, number of surveillance cameras and the purpose of the dataset.

Dataset	Category	Input format	Number of identical vehicles	Mode of data acquisition	# Surveillance platform (CCTVs/ UAVs)	Relevance
CompCars [190]	Cross modal	Image	1,716	Web and Survielance Cameras	-	Re-identification
VehicleID [8]	Single Platform	Image	26,267	Surveillance cameras	Multiple Cameras	Re-identification
VeRi-776 [40]	Single Platform	Image	776	Surveillance cameras	20 CCTVs	Re-identification
VERI-Wild [51]	Single Platform	Image	40,671	Surveillance cameras	174 CCTVs	Re-identification
VERI-Wild 2.0 [194]	Single Platform	Image	42,790	Surveillance cameras	274 CCTVs	Re-identification
VehicleX [41]	Single Platform	Synthetic Images	1,362	3D generated	-	Re-identification, Semantic Segmentation, Object detection and 3D reconstruction
MCU-VReID [13]	Cross Platform	Image	51	CCTVs and UAV	42 CCTVs and UAV	Re-identification
CityFlow [104]	Single Platform	Videos	-	CCTVs	40 CCTVs	Re-identification, MTMC tracking
RGBN300 [89]	Multi-modal	Multi-spectral images	300	8 Cameras	-	Re-identification
MSVR310 [93]	Multi-modal	Multi-spectral images	310	D866 Camera, FLIR SC620 camera and Mi8 mobile phone camera	-	Re-identification
VRAI [11]	Single Platform	Image	13,022	Aerial Devices	2 UAVs	Re-identification
UAV-VeID [12]	Single Platform	Image	4,601	Aerial Devices	2 UAVs	Re-identification
VRU [195]	Single Platform	Image	15,085	Aerial Devices	5 UAVs	Re-identification
Synthehicle [181]	Single Platform	Synthetic Images	-	Surveillance cameras	16 CCTVs	Vehicle Tracking, 2D & 3D detection and tracking, depth estimation, and semantic, instance and panoptic segmentation.
UAV123 [183]	Single Platform	UAV Videos and Synthetic Videos	-	Aerial Devices and UAV Simulator	1 UAV	Tracking, detection and Segmentation
UAVDT [189]	Single Platform	Videos	-	Aerial Device	1 UAV	Tracking, detection and Segmentation

Moreover, the physical environment poses obstacles such as variable illumination and swiftly moving vehicles, thereby impeding the consistent capture of superior images. In addition to these limitations, the class of vehicles may exhibit an imbalance, where commonly observed vehicle models are significantly more prevalent compared to rare ones. This disparity can result in the system facing difficulties in accurately recognizing infrequent vehicle categories. Creating large and diverse datasets for training purposes becomes costly due to the substantial amount of time and resources needed to manually label vehicle images. The lack of available data can greatly affect the accuracy and reliability of vehicle re-identification systems.

B. SUBTLE VARIATION IN INTRA AND INTER CLASS VEHICLE INSTANCES

The task of vehicle re-identification involves finding the best possible match for a queried vehicle across neighboring cameras. Unlike person re-identification, where the semantic structure of a person remains relatively constant, vehicles exhibit different semantic structures when viewed from various perspectives. This variation makes it challenging to identify identical vehicles that have similar appearances but belong to different model categories. Intra-class variations present a significant challenge in cross platform vehicle re-identification, especially when vehicles are observed from both aerial and ground-level views. These differing perspectives complicate the inference of intra-vehicle similarities, potentially leading to a decline in re-identification accuracy.

C. HANDLING DIFFERENT FORMS OF INPUTS

In recent years, significant efforts have been made to tackle the challenges of vehicle re-identification in low-light conditions, night time scenes, foggy weather, and other challenging environments. Various multi-spectral image acquisition techniques have been employed to address these issues. However, these techniques necessitate additional investment and manpower for data collection, and the loss of information from any one image format can significantly degrade re-identification performance. Moreover, designing a re-identification framework that effectively utilizes these diverse image forms requires carefully designed feature extractors, which can be resource-intensive and demand substantial storage capacity. Recent studies [196] have explored the use of complementary inputs, such as images along with natural language descriptions, to enhance re-identification accuracy. This form of re-identifying vehicles refers to cross-modality vehicle re-identification that aims to re-identify and match vehicles across different modalities, such as images and textual descriptions. This approach leverages attributes and textual descriptions to establish connections between visual representations of vehicles and their corresponding textual features. In cross-modal re-identification, the re-identification algorithm is pre-trained on a diverse dataset, so that the model can learn generalized representations that enhance its ability to identify vehicles even in unseen re-identification datasets. The inference process involves utilizing the learned cross-modal features to match vehicle images with their corresponding textual descriptions in the unseen dataset, thereby facilitating effective identification despite variability in appearance or context. Integrating these different machine learning principles poses its own set of challenges, but offers promising improvements in re-identification performance.

D. SCALE, EGO MOTION, CHANGING SCENES IN UAVS

The re-identification and tracking of vehicles using UAVs encounter various challenges associated with scale, ego motion, and dynamic scenes. The different elevations at which UAVs operate lead to substantial differences in the size of the captured vehicles, making it challenging to consistently identify them. Ego motion, which refers to the movement of the UAV itself, adds additional complexity as it causes the camera's viewpoint to constantly change. This results in varying perspectives and the possibility of motion blur. To achieve accurate tracking, it is necessary to use strong algorithms that can make real-time adjustments in response to the dynamic movement. Moreover, the imagery obtained by UAVs undergoes constant variations because of environmental conditions, traffic patterns, and diverse backgrounds. These factors have the potential to obscure or modify the visual characteristics of vehicles. To address these challenges, it is crucial to develop advanced re-identification frameworks that can effectively handle scale invariance, compensate for ego motion, and adapt to various scene variations.

E. ASSOCIATION OF CROSS PLATFORM VEHICLE TRACKLETS

Cross platform vehicle re-identification, recently introduced to track vehicles across different surveillance platforms, presents significant challenges. One key issue in cross platform vehicle tracking is the dynamic nature of scenes captured by UAVs. As UAVs move, the scenes change rapidly, while only a limited number of ground-level surveillance cameras capture the same vehicles observed by the UAVs. The varying motion of UAVs makes it difficult to associate vehicle tracklets from aerial footage with those from ground cameras. This inconsistency leads to fragmented vehicle tracklets for the same vehicle, often resulting in incorrect assignment of tracklet IDs. Addressing these challenges is crucial for achieving accurate and reliable cross platform vehicle re-identification.

Despite the challenges associated with vehicle re-identification and tracking, researchers can explore several potential future directions to address these issues. By considering advanced methodologies and innovative approaches, it is possible to overcome the current hurdles and improve the accuracy and reliability of re-identification and tracking systems. The study presents possible directions that could be considered in future for designing a re-identification datasets and frameworks.

F. EXPLORING DIFFERENT TRAINING STRATEGIES

While designing the re-identification framework where there is scarcity of labeled data, it is feasible to look for different training strategies to learn and adapt the re-identification frame-work to learn more robust features of vehicles. Different training strategies such as unsupervised, semi-supervised and self-supervised learning can be looked upon for training the re-identification networks. Techniques such as teacher student model, knowledge distillation, and a few shot learning could be explored when there is scarcity of target labeled data. Additionally, for scenarios where different data modalities like images and textual descriptions are available, adopting a cross modal learning approach can be beneficial. By adaptively incorporating these different modalities, the model can achieve better recognition even with limited data.

G. 3D RE-IDENTIFICATION

Developments in 3D vehicle re-identification have great possibility to solve present constraints in 2D approaches such as perspective variance, occlusions, and subtle inter-class discrepancies [197], [198]. Leveraging 3D modelling, methods such perspective alignment, 3D mesh generation, and depth-based feature enhancement can yield strong representations that are invariant to camera distortions and dynamic surroundings [197]. Particularly for rigid and deformable objects, including point cloud data from LiDAR and other depth sensors provides an extra degree of accuracy, hence enabling detailed shape-based re-identification [199]. Recent efforts show cooperative frameworks merging vehicle and infrastructure side 3D perceptions suggesting a scalable future for real-time, cross-domain re-identification [200]. By allowing exact tracking across several viewpoints and platforms, such techniques can improve autonomous driving, urban traffic surveillance, and smart city systems [197], [200]. Future work can concentrate on enhancing computing efficiency, data fusion techniques, and lightweight 3D re-identification models to enable this strategy for aiding re-identification in traffic surveillance for ITS.

H. RE-IDENTIFICATION UNDER EXTREME WEATHER CONDITIONS

Inclement weather conditions, including precipitation, snowfall, fog, and poor lighting, substantially diminish the efficacy of vehicle re-identification algorithms by compromising image quality and contrast, thereby hindering feature extraction. These factors provide further hurdles such as occlusions, change in illumination, and noise, which impede the model's capacity to uphold consistent representations across images. To mitigate these challenges, effective re-identification systems frequently include weather-invariant feature learning, domain adaptation methodologies, or data augmentation techniques that replicate various weather conditions during training. Re-identification problems could be modelled using techniques such as Blind Image Decomposition [201] to incorporate depth information for image restoration during adverse weather conditions. Additionally, the methodologies described in [202] could be implemented while re-identifying vehicles using UAVs to mitigate domain shifts that result from fluctuating environmental conditions.

I. HYBRID APPROACH FOR VEHICLE RE-IDENTIFICATION AND TRACKING

The fusion of UAVs, CCTVs and multispectral imaging (IR, thermal, and RGB) can greatly improve vehicle re-identification. High resolution ground-level surveillance is offered by CCTVs, which may record important vehicle characteristics like license plates and colors. UAVs provide airborne views, capturing larger and more challenging to reach areas. This data is enhanced by multispectral imaging: While IR photos identify heat from recently driven vehicles and TIR images show persistent thermal features, RGB photographs capture color and texture. Contextual information that images might miss is provided by adding written descriptions from traffic reports or witnesses. According to the idea, integrating these technologies results in a strong system for monitoring vehicles that takes use of each one's advantages for increased precision and fewer false positives. The goal of this hybrid strategy is to improve vehicle tracking in a variety of settings and circumstances.

J. CONTEXTUAL INFORMATION FOR ASSOCIATING VEHICLE TRACKLETS

Contextual information provides an extra degree of precision to vehicle tracklet association, even if appearance and

temporal aspects are still important. This context can include nighttime illumination variations that could modify the look of a vehicle, or lane changes detected by a UAV that fills a gap in a CCTV track. This context may be extracted by computer vision and deep learning techniques. For example, scene understanding techniques using semantic segmentation may examine lighting conditions, while object detection algorithms can recognize traffic lights and lane markings. Through the integration of contextual cues alongside traditional features, the system can establish stronger linkages between tracklets, even in the presence of occlusions, difficult lighting conditions, or transient changes in appearance.

VI. CONCLUSION

This paper discusses recent advancements in the field of vehicle re-identification and tracking, emphasizing their importance for intelligent transportation systems (ITS). Vehicle re-identification tasks are categorized into single platform, multi-spectral/multi-modal, cross domain, and cross platform re-identification. Additionally, the study highlights various vehicle tracking frameworks, primarily those presented in the AI City Challenge and other datasets utilizing UAVs. The discussion extends to the datasets available for each category of re-identification and tracking. The paper identifies certain drawbacks that still need to be addressed to improve re-identification and tracking performance. Finally, it outlines the future scope for designing and enhancing these systems to implement better security measures and enhance safety in ITS.

APPENDIX

Search Strategy for Papers of Vehicle Re-Identification and Tracking: This review article primarily draws upon research papers obtained from the Scopus database. The articles were identified using an "Advanced Search" query with the keyword "Vehicle re-identification". To focus on recent advancements, the search was refined to include works published between 2019 and 2024, resulting in an initial collection of 890 articles. These were further narrowed down based on specific inclusion criteria: the "Document Type" was restricted to Articles, Conference Papers, and Reviews; the "Keyword" was limited to Vehicle re-identification; the "Source Type" was confined to Journals and Conference Proceedings; and only papers in English were considered. This filtering yielded a total of 281 articles. Finally, articles unavailable under "View at Publisher" or those lacking validation or evaluation on benchmark datasets listed in Table 3 were excluded. The table is provided in Supplementary material.

REFERENCES

[1] S. K. Jagatheesaperumal, S. E. Bibri, J. Huang, J. Rajapandian, and B. Parthiban, "Artificial intelligence of things for smart cities: Advanced solutions for enhancing transportation safety," *Comput. Urban Sci.*, vol. 4, no. 1, p. 10, 2024, doi: 10.1007/s43762-024-00120-6.

[2] Q. Cheng, Y. Wang, W. He, and Y. Bai, "Lightweight air-to-air unmanned aerial vehicle target detection model," *Sci. Rep.*, vol. 14, no. 1, pp. 1–18, 2024, doi: 10.1038/s41598-024-53181-2.

[3] J. Prakash, L. Murali, N. Manikandan, N. Nagaprasad, and K. Ramaswamy, "A vehicular network based intelligent transport system for smart cities using machine learning algorithms," *Sci. Rep.*, vol. 14, no. 1, p. 468, 2024.

[4] E. Tian and J. Kim, "De-hazing CCTV images using dark channel prior for improved vehicle detection," in *Proc. 8th Int. Conf. Intell. Inf. Technol. (ICIIT)*, 2023, pp. 1152–156. [Online]. Available: https://doi.org/10.1145/3591569.3591597

[5] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "On-road vehicle detection and tracking using MMW radar and monovision fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 2075–2084, Jul. 2016, doi: 10.1109/TITS.2016.2533542.

[6] B. Hardjono, "Vehicle counting tool interface design for machine learning methods," in *Proc. 11th Int. Conf. Inf. Technol. IoT Smart City (ICIT)*, 2024, pp. 1–9. [Online]. Available: https://doi.org/10.1145/3638985.3638986

[7] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1635–1647, Mar. 2014, doi: 10.1109/TGRS.2013.2253108.

[8] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 869–884.

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

[10] D. Song, R. Tharmarasa, T. Kirubarajan, and X. N. Fernando, "Multi-vehicle tracking with road maps and car-following models," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1375–1386, May 2018, doi: 10.1109/TITS.2017.2723575.

[11] P. Wang et al., "Vehicle re-identification in aerial imagery: Dataset and approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 460–469.

[12] S. Teng, S. Zhang, Q. Huang, and N. Sebe, "Viewpoint and scale consistency reinforcement for UAV vehicle re-identification," *Int. J. Comput. Vis.*, vol. 129, pp. 719–735, Mar. 2021, doi: 10.1007/s11263-020-01402-2.

[13] B. A. Holla, M. M. M. Pai, U. Verma, and R. M. Pai, "MSFFT: Multi-scale feature fusion transformer for cross platform vehicle re-identification," *Neurocomputing*, vol. 582, May 2024, Art. no. 127514. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224002856

[14] R. D. Kuhne and S. Immes, "Freeway control systems for using section-related traffic variable detection," in *Proc. Pac. Rim TransTech Conf.*. vol. 1, 1993, p. 234.

[15] L. Klein, M. Kelley, and M. Mills, "Evaluation of overhead and in-ground vehicle detector technologies for traffic flow measurement," *J. Test. Eval.*, vol. 25, no. 2, pp. 205–214, Mar. 1997. [Online]. Available: https://doi.org/10.1520/JTE11480J

[16] K. N. Balke, G. Ullman, W. McCasland, C. Mountain, and C. Dudek, "Benefits of real-time travel information in Houston, Texas," Texas Transp. Inst., Arlington, TX, USA, Rep. SWUTC/95/60010-1, 1995.

[17] Z. Tang et al., "Multiple-kernel based vehicle tracking using 3-D deformable model and camera self-calibration," 2017, *arXiv:1708.06831*.

[18] T. Liu and Y. Liu, "Deformable model-based vehicle tracking and recognition using 3-D constrained multiple-kernels and Kalman filter," *IEEE Access*, vol. 9, pp. 90346–90357, 2021.

[19] A. Ayala-Acevedo, A. Devgun, S. Zahir, and S. Askary, "Vehicle re-identification: Pushing the limits of re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 291–296.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: https://doi.org/10.1145/3065386

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[22] S. H. Silva, P. Rad, N. Beebe, K.-K. R. Choo, and M. Umapathy, "Cooperative unmanned aerial vehicles with privacy preserving deep vision for real-time object identification and tracking," *J. Parallel Distrib. Comput.*, vol. 131, pp. 147–160, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0743731518308839

[23] G. Cho, Y. Shinyama, J. Nakazato, K. Maruta, and K. Sakaguchi, "Object recognition network using continuous roadside cameras," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC-Spring)*, 2022, pp. 1–5.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[26] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Red Hook, NY, USA: Curran Assoc., 2019.

[27] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731421930030X

[28] 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.

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

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2261–2269.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

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

[33] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[34] J. Yu, H. Oh, M. Kim, and J. Kim, "Weakly supervised contrastive learning for unsupervised vehicle reidentification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15543–15553, Nov. 2024, doi: 10.1109/TNNLS.2023.3288139.

[35] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3908–3916.

[36] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1, pp. 1–20, Dec. 2017. [Online]. Available: https://doi.org/10.1145/3159171

[37] H. Chen et al., "Deep transfer learning for person re-identification," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, 2018, pp. 1–5.

[38] Z. Sun, X. Nie, X. Xi, and Y. Yin, "CFVMNET: A multi-branch network for vehicle re-identification based on common field of view," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, 2020, pp. 3523–3531. [Online]. Available: https://doi.org/10.1145/3394171.3413541

[39] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3690–3700.

[40] 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)*, 2016, pp. 1–6.

[41] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon, "Simulating content consistent vehicle datasets with attribute descent," in *Proc. Comput. Vis. (ECCV)*, 2020, pp. 775–791.

[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6629–6640.

[43] T.-S. Chen, M.-Y. Lee, C.-T. Liu, and S.-Y. Chien, "Viewpoint-aware channel-wise attentive network for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 2448–2455.

[44] M.-C. Chang et al., "AI city challenge 2020—Computer vision for smart transportation applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 2638–2647.

[45] 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)*, 2021, pp. 14993–15002.

[46] S. Teng, S. Zhang, Q. Huang, and N. Sebe, "Multi-view spatial attention embedding for vehicle re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 816–827, Feb. 2021, doi: 10.1109/TCSVT.2020.2980283.

[47] H. Li et al., "Attributes guided feature learning for vehicle re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 5, pp. 1211–1221, Oct. 2022, doi: 10.1109/TETCI.2021.3127906.

[48] Q. Wang et al., "Viewpoint adaptation learning with cross-view distance metric for robust vehicle re-identification," *Inf. Sci.*, vol. 564, pp. 71–84, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521001559

[49] H. Wang, J. Peng, G. Jiang, F. Xu, and X. Fu, "Discriminative feature and dictionary learning with part-aware model for vehicle re-identification," *Neurocomputing*, vol. 438, pp. 55–62, May 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122100117X

[50] S. Zhang, C. Lin, and S. Ma, "Large margin metric learning for multi-view vehicle re-identification," *Neurocomputing*, vol. 447, pp. 118–128, Aug. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221004288

[51] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3230–3238.

[52] J. Peng, G. Jiang, and H. Wang, "Generalized multiple sparse information fusion for vehicle re-identification," *J. Vis. Commun. Image Represent.*, vol. 79, Aug. 2021, Art. no. 103207. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320321001334

[53] O. Moskvyak, F. Maire, F. Dayoub, and M. Baktashmotlagh, "Keypoint-aligned embeddings for image retrieval and re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 676–685.

[54] R. Kuma, E. Weill, F. Aghdasi, and P. Sriram, "Vehicle re-identification: An efficient baseline using triplet embedding," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–9.

[55] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1005–1014.

[56] X. Fu, J. Peng, G. Jiang, and H. Wang, "Learning latent features with local channel drop network for vehicle re-identification," *Eng. Appl. Artif. Intell.*, vol. 107, Jan. 2022, Art. no. 104540. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197621003882

[57] B. A. Holla, 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.

[58] Y. Zheng, X. Pang, G. Jiang, X. Tian, and Q. Meng, "Dual-relational attention network for vehicle re-identification," *Appl. Intell.*, vol. 53, no. 7, pp. 7776–7787, Jul. 2022. [Online]. Available: https://doi.org/10.1007/s10489-022-03801-z

[59] B. Li et al., "VehicleGAN: Pair-flexible pose guided image synthesis for vehicle re-identification," in *Proc. IEEE Intell. Veh. Symp. (IV)*, 2024, pp. 447–453.

[60] A. Ayub and H. Kim, "GAN-based data augmentation with vehicle color changes to train a vehicle detection CNN," *Electronics*, vol. 13, no. 7, p. 1231, 2024.

[61] F. Zhang, Y. Ma, G. Yuan, H. Zhang, and J. Ren, "Multiview image generation for vehicle reidentification," *Appl. Intell.*, vol. 51, no. 8, pp. 5665–5682, 2021.

[62] Y. Xie, H. Wu, J. Zhu, and H. Zeng, "Distillation embedded absorbable pruning for fast object re-identification," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110437. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320324001882

[63] Y. Xie, H. Wu, Y. Lin, J. Zhu, and H. Zeng, "Pairwise difference relational distillation for object re-identification," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110455. [Online]. Available: https://doi.org/10.1016/j.patcog.2024.110455

[64] X. Pang, Y. Zheng, X. Nie, Y. Yin, and X. Li, "Multi-axis interactive multidimensional attention network for vehicle re-identification," *Image Vis. Comput.*, vol. 144, Apr. 2024, Art. no. 104972. [Online]. Available: https://doi.org/10.1016/j.imavis.2024.104972

[65] W. Hu, H. Zhan, P. Shivakumara, U. Pal, and Y. Lu, "TANET: Text region attention learning for vehicle re-identification," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108448. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624006067

[66] R. Kishore, N. Aslam, and M. H. Kolekar, "PATReId: Pose apprise transformer network for vehicle re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 5, pp. 3691–3702, Oct. 2024, doi: 10.1109/TETCI.2024.3372391.

[67] B. Jiao, L. Yang, L. Gao, P. Wang, S. Zhang, and Y. Zhang, "Vehicle re-identification in aerial images and videos: Dataset and approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1586–1603, Mar. 2024, doi: 10.1109/TCSVT.2023.3298788.

[68] S. Chen, M. Ye, and B. Du, "Rotation invariant transformer for recognizing object in UAVs," in *Proc. 30th ACM Int. Conf. Multimedia (MM)*, 2022, pp. 2565–2574. [Online]. Available: https://doi.org/10.1145/3503161.3547799

[69] C. Zhang, C. Yang, D. Wu, H. Dong, and B. Deng, "Cross-view vehicle re-identification based on graph matching," *Appl. Intell.*, vol. 52, no. 13, pp. 14799–14810, Oct. 2022. [Online]. Available: https://doi.org/10.1007/s10489-022-03349-y

[70] A. Yao, J. Qi, and P. Zhong, "Self-aligned spatial feature extraction network for UAV vehicle reidentification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[71] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[72] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," 2021, *arXiv:2104.02008*.

[73] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 1487–1495.

[74] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3140–3149.

[75] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model HETERO-and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–188.

[76] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 598–607.

[77] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.

[78] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2242–2251.

[79] D. Fu et al., "Unsupervised pre-training for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14745–14754.

[80] J. Peng, H. Wang, F. Xu, and X. Fu, "Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification," *Neurocomputing*, vol. 401, pp. 133–144, Aug. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122030357X

[81] H. Zhang, Z. Kuang, L. Cheng, Y. Liu, X. Ding, and Y. Huang, "AIVR-Net: Attribute-based invariant visual representation learning for vehicle re-identification," *Knowl.-Based Syst.*, vol. 289, Apr. 2024, Art. no. 111455. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070512400090X

[82] Y. Li, F. Yang, Y. Tian, X. Wang, Q. Chen, and P. Jing, "Collaborative representation based cross-domain semantic transfer for vehicle re-identification," *Neurocomputing*, vol. 567, Jan. 2024, Art. no. 127039. [Online]. Available: https://doi.org/10.1016/j.neucom.2023.127039

[83] J. Peng, H. Wang, T. Zhao, and X. Fu, "Cross domain knowledge transfer for unsupervised vehicle re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2019, pp. 453–458.

[84] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1857–1865.

[85] Y. Wang, J. Peng, H. Wang, and M. Wang, "Progressive learning with multi-scale attention network for cross-domain vehicle re-identification," *Sci. China Inf. Sci.*, vol. 65, no. 6, 2022, Art. no. 160103.

[86] L. Song et al., "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.107173

[87] R. Bashir, M. Shahzad, and M. Fraz, "VR-Proud: Vehicle re-identification using progressive unsupervised deep architecture," *Pattern Recognit.*, vol. 90, pp. 52–65, Jun. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319300147

[88] J. Peng, Y. Wang, H. Wang, Z. Zhang, X. Fu, and M. Wang, "Unsupervised vehicle re-identification with progressive adaptation," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2020, pp. 913–919. [Online]. Available: https://doi.org/10.24963/ijcai.2020/127

[89] H. Li, C. Li, X. Zhu, A. Zheng, and B. Luo, "Multi-spectral vehicle re-identification: A challenge," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 11345–11353. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6796

[90] Q. He, Z. Lu, Z. Wang, and H. Hu, "Graph-based progressive fusion network for multi-modality vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12431–12447, Nov. 2023, doi: 10.1109/TITS.2023.3285758.

[91] W. Pan, H. Wu, J. Zhu, H. Zeng, and X. Zhu, "H-VIT: Hybrid vision transformer for multi-modal vehicle re-identification," in *Proc. 2nd CAAI Int. Conf. Artif. Intell. (CICAI)*, vol. 13604, 2022, pp. 255–267. [Online]. Available: https://doi.org/10.1007/978-3-031-20497-5_21

[92] J. Guo, X. Zhang, Z. Liu, and Y. Wang, "Generative and attentive fusion for multi-spectral vehicle re-identification," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, 2022, pp. 1565–1572.

[93] A. Zheng, X. Zhu, Z. Ma, C. Li, J. Tang, and J. Ma, "Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101901. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523002178

[94] A. Zheng, Z. He, Z. Wang, C. Li, and J. Tang, "Dynamic enhancement network for partial multi-modality person re-identification," 2023, *arXiv:2305.15762*.

[95] H. Yin, J. Li, E. Schiller, L. McDermott, and D. Cummings, "GRAFT: Gradual fusion transformer for multimodal re-identification," 2023, *arXiv:2310.16856*.

[96] J. Crawford, H. Yin, L. McDermott, and D. Cummings, "Unicat: Crafting a stronger fusion baseline for multimodal re-identification," 2023, *arXiv:2310.18812*.

[97] E. Kamenou, J. M. D. Rincón, P. Miller, and P. Devlin-Hill, "A meta-learning approach for domain generalisation across visual modalities in vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2023, pp. 385–393.

[98] X. Wang, X. Wang, B. Jiang, J. Tang, and B. Luo, "MutualFormer: Multi-modal representation learning via cross-diffusion attention," *Int. J. Comput. Vis.*, vol. 132, no. 9, pp. 3867–3888, 2024.

[99] H. Liu, X. Tan, and X. Zhou, "Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 4414–4425, 2021, doi: 10.1109/TMM.2020.3042080.

[100] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 618–626.

[101] Z. Wang, C. Li, A. Zheng, R. He, and J. Tang, "Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 2633–2641. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20165

[102] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3529–3537, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16467

[103] B. A. Holla, M. M. Pai, U. Verma, and R. M. Pai, "Vehicle re-identification in smart city transportation using hybrid surveillance systems," in *Proc. IEEE Region 10 Conf. (TENCON)*, 2021, pp. 335–340.

[104] Z. Tang et al., "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 8789–8798.

[105] B. He, J. Li, Y. Zhao, and Y. Tian, "Part-regularized near-duplicate vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3992–4000.

[106] Z. Tang et al., "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 211–220.

[107] R. Chu, Y. Sun, Y. Li, Z. Liu, C. Zhang, and Y. Wei, "Vehicle re-identification with viewpoint-aware metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8282–8291.

[108] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A dual-path model with adaptive attention for vehicle re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6131–6140.

[109] X. Liu, S. Zhang, X. Wang, R. Hong, and Q. Tian, "Group-group loss-based global-regional feature learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 2638–2652, 2020.

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

[111] D. Meng et al., "Parsing-based view-aware embedding network for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7101–7110.

[112] T.-S. Chen, C.-T. Liu, C.-W. Wu, and S.-Y. Chien, "Orientation-aware vehicle re-identification with semantics-guided part attention network," in *Proc. Comput. Vis. 16th Eur. Conf. (ECCV)*, Aug. 2020, pp. 330–346.

[113] P. Khorramshahi, N. Peri, J.-C. Chen, and R. Chellappa, "The devil is in the details: Self-supervised attention for vehicle re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 369–386.

[114] 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*, 2020, pp. 907–915.

[115] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "VehicleNet: Learning robust visual representation for vehicle re-identification," *IEEE Trans. Multimedia*, vol. 23, pp. 2683–2693, 2020.

[116] W. Sun, G. Dai, X. Zhang, X. He, and X. Chen, "TBE-Net: A three-branch embedding network with part-aware ability and feature complementary learning for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14557–14569, Sep. 2022.

[117] J. Zhao, F. Qi, G. Ren, and L. Xu, "PHD learning: Learning with Pompeiu–Hausdorff distances for video-based vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2225–2235.

[118] X. Zhang, Y. Ge, Y. Qiao, and H. Li, "Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 3435–3444.

[119] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4682–4692.

[120] A. Wu, W. Ge, and W.-S. Zheng, "Rewarded semi-supervised re-identification on identities rarely crossing camera views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15512–15529, Dec. 2023.

[121] X. Zhou, Y. Zhong, Z. Cheng, F. Liang, and L. Ma, "Adaptive sparse pairwise loss for object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19691–19701.

[122] Y. Yao, T. Gedeon, and L. Zheng, "Large-scale training data search for object re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15568–15578.

[123] W. Pan, L. Huang, J. Liang, L. Hong, and J. Zhu, "Progressively hybrid transformer for multi-modal vehicle re-identification," *Sensors*, vol. 23, no. 9, p. 4206, 2023.

[124] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, May 2008.

[125] R. D. Iaco, S. L. Smith, and K. Czarnecki, "Universally safe swerve maneuvers for autonomous driving," *IEEE Open J. Intell. Transp. Syst.*, vol. 2, pp. 482–494, 2021, doi: 10.1109/OJITS.2021.3138953.

[126] J. Betz et al., "Autonomous vehicles on the edge: A survey on autonomous vehicle racing," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 458–488, 2022, doi: 10.1109/OJITS.2022.3181510.

[127] C.-L. Lee, C.-Y. Hou, C.-C. Wang, and W.-C. Lin, "Extrinsic and temporal calibration of automotive radar and 3-D LiDAR in factory and on-road calibration settings," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 708–719, 2023, doi: 10.1109/OJITS.2023.3312660.

[128] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006, doi: 10.1109/TPAMI.2006.104.

[129] Y. Qian, X. Wang, H. Zhuang, C. Wang, and M. Yang, "3-D vehicle detection enhancement using tracking feedback in sparse point clouds environments," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 471–480, 2023, doi: 10.1109/OJITS.2023.3283768.

[130] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec. 2013, doi: 10.1109/TITS.2013.2266661.

[131] F. Leon and M. Gavrilescu, "A review of tracking and trajectory prediction methods for autonomous driving," *Mathematics*, vol. 9, no. 6, p. 660, 2021.

[132] L. Rakai, H. Song, S. Sun, W. Zhang, and Y. Yang, "Data association in multiple object tracking: A survey of recent techniques," *Exp. Syst. Appl.*, vol. 192, May 2022, Art. no. 116300.

[133] V. Kamath and A. Renuka, "Deep learning based object detection for resource constrained devices: Systematic review, future trends and challenges ahead," *Neurocomputing*, vol. 531, pp. 34–60, Apr. 2023.

[134] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, CNNs and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023.

[135] B. Kaur and S. Singh, "Object detection using deep learning: A review," in *Proc. Int. Conf. Data Sci. Mach. Learn. Artif. Intell.*, 2021, pp. 328–334.

[136] Y. Xiao et al., "A review of object detection based on deep learning," *Multimedia Tools Appl.*, vol. 79, no. 1, pp. 23729–23791, 2020.

[137] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digit. Signal Process.*, vol. 132, 2023, Art. no. 103812.

[138] M. Naphade et al., "The 5th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4258–4268.

[139] M. Naphade et al., "The 6th AI city challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3346–3355.

[140] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann, "Electricity: An efficient multi-camera vehicle tracking system for intelligent city," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 2511–2519.

[141] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[142] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 3645–3649.
[143] N. Peri et al., "Towards real-time systems for vehicle re-identification, multi-camera tracking, and anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 2648–2657.
[144] Z. Tang and J.-N. Hwang, "MOANA: An online learned adaptive appearance model for robust multiple object tracking in 3-D," *IEEE Access*, vol. 7, pp. 31934–31945, 2019.
[145] K.-H. N. Bui, H. Yi, and J. Cho, "A vehicle counts by class framework using distinguished regions tracking at multiple intersections," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 2466–2474.
[146] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.
[147] Z. Wang et al., "Robust and fast vehicle turn-counts at intersections via an integrated solution from detection, tracking and trajectory modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 2598–2606.
[148] C. Liu et al., "City-scale multi-camera vehicle tracking guided by crossroad zones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4124–4132.
[149] M. Wu, Y. Qian, C. Wang, and M. Yang, "A multi-camera vehicle tracking system based on city-scale vehicle RE-ID and spatial–temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4072–4081.
[150] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection ," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01079
[151] Y.-L. Li, Z.-Y. Chin, M.-C. Chang, and C.-K. Chiang, "Multi-camera tracking by candidate intersection ratio tracklet matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4098–4106.
[152] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, May 2017.
[153] P. Ren et al., "Multi-camera vehicle tracking system based on spatial–temporal filtering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4208–4214.
[154] K. Shim, S. Yoon, K. Ko, and C. Kim, "Multi-target multi-camera vehicle tracking for city-scale traffic management," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4188–4195.
[155] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5987–5995.
[156] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.
[157] A. Specker, D. Stadler, L. Florin, and J. Beyerer, "An occlusion-aware multi-target multi-camera tracking system," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4173–4182.
[158] K.-S. Yang, Y.-K. Chen, T.-S. Chen, C.-T. Liu, and S.-Y. Chien, "Tracklet-refined multi-camera tracking based on balanced cross-domain re-identification for vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 3978–3987.
[159] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with trackletnet," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, 2019, pp. 482–490. [Online]. Available: https://doi.org/10.1145/3343031.3350853
[160] J. Ye et al., "A robust MTMC tracking system for AI-city challenge 2021," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4039–4048.
[161] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[162] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.
[163] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.
[164] H. Yao et al., "City-scale multi-camera vehicle tracking based on space–time–appearance features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3309–3317.
[165] A. Specker, L. Florin, M. Cormier, and J. Beyerer, "Improving multi-target multi-camera tracking by track refinement and completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3198–3208.
[166] F. Li et al., "Multi-camera vehicle tracking system for AI city challenge 2022," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3264–3272.
[167] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2016, pp. 3464–3468.
[168] N. M. Chung et al., "Multi-camera multi-vehicle tracking with domain generalization and contextual constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3326–3336.
[169] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6931–6939.
[170] Y. Liu, X. Zhang, B. Zhang, X. Zhang, S. Wang, and J. Xu, "Multi-camera vehicle tracking based on occlusion-aware and inter-vehicle information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3256–3263.
[171] X. Yang et al., "Box-grained reranking matching for multi-camera multi-target tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2022, pp. 3095–3105.
[172] H.-M. Hsu, Y. Wang, and J.-N. Hwang, "Traffic-aware multi-camera tracking of vehicles based on REID and camera link model," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 964–972.
[173] Y. Liu, L. Zhang, Z. Chen, Y. Yan, and H. Wang, "Multi-stream Siamese and faster region-based neural network for real-time object tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7279–7292, Nov. 2021, Doi: 10.1109/TITS.2020.3006927.
[174] H.-M. Hsu, J. Cai, Y. Wang, J.-N. Hwang, and K.-J. Kim, "Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model," *IEEE Trans. Image Process.*, vol. 30, pp. 5198–5210, 2021, doi: 10.1109/TIP.2021.3078124.
[175] G. Wang, R. Gu, Z. Liu, W. Hu, M. Song, and J.-N. Hwang, "Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking," in textitProc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9876–9886.
[176] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[177] L. Wen et al., "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understand.*, vol. 193, Jan. 2020, Art. no. 102907.
[178] H.-N. Hu, Y.-H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun, "Monocular quasi-dense 3-D object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1992–2008, Feb. 2023, doi: 10.1109/TPAMI.2022.3168781.
[179] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.
[180] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2446–2454.
[181] F. Herzog, J. Chen, T. Teepe, J. Gilg, S. Hörmann, and G. Rigoll, "Synthehicle: Multi-vehicle multi-camera tracking in virtual cities," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1–11.

[182] P. Nousi, D. Triantafyllidou, A. Tefas, and I. Pitas, "Re-identification framework for long term visual object tracking based on object detection and classification," *Signal Process. Image Commun.*, vol. 88, Oct. 2020, Art. no. 115969.
[183] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 445–461.
[184] A. Moudgil and V. Gandhi, "Long-term visual object tracking benchmark," in *Proc. Comput. Vis. (ACCV) 14th Asian Conf. Comput. Vis.*, Dec. 2019, pp. 629–645.
[185] C. Lusardi, A. M. N. Taufique, and A. Savakis, "Robust multi-object tracking using re-identification features and graph convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3868–3877.
[186] A. M. N. Taufique and A. Savakis, "LabNet: Local graph aggregation network with class balanced loss for vehicle re-identification," *Neurocomputing*, vol. 463, pp. 122–132, Nov. 2021.
[187] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 3069–3087, Sep. 2021.
[188] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[189] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 370–386.
[190] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3973–3981.
[191] S. Girisha, M. M. M. Pai, U. Verma, and R. M. Pai, "Performance analysis of semantic segmentation algorithms for finely annotated new UAV aerial video dataset (ManipalUAVid)," *IEEE Access*, vol. 7, pp. 136239–136253, 2019.
[192] S. Girisha, U. Verma, M. M. M. Pai, and R. M. Pai, "UVID-Net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4115–4127, 2021.
[193] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learn.*, 2017, pp. 1–16.
[194] 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.
[195] M. Lu, Y. Xu, and H. Li, "Vehicle re-identification based on UAV viewpoint: Dataset and method," *Remote Sens.*, vol. 14, no. 18, p. 4603, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/18/4603
[196] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 4492–4501.
[197] D. Meng, L. Li, X. Liu, L. Gao, and Q. Huang, "Viewpoint alignment and discriminative parts enhancement in 3-D space for vehicle REID," *IEEE Trans. Multimedia*, vol. 25, pp. 2954–2965, 2023, doi: 10.1109/TMM.2022.3154102.
[198] T.-W. Huang, J. Cai, H. Yang, H.-M. Hsu, and J.-N. Hwang, "Multi-view vehicle re-identification using temporal attention model and metadata re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 434–442.
[199] B. Thérien, C. Huang, A. Chow, and K. Czarnecki, "Object re-identification from point clouds," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 8377–8388.
[200] C. Sun, Y. Wang, H. Li, J. Guo, and Y. Deng, "Collaborative and reidentifying techniques for improved monocular 3-D perception in vehicles," *IEEE Internet Things J.*, vol. 11, no. 22, pp. 36358–36369, Nov. 2024.
[201] C. Wang, Z. Zheng, R. Quan, and Y. Yang, "Depth-aware blind image decomposition for real-world adverse weather recovery," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 379–397.
[202] T. Wang, Z. Zheng, Y. Sun, C. Yan, Y. Yang, and T.-S. Chua, "Multiple-environment self-adaptive network for aerial-view geo-localization," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110363. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320324001146

ASHUTOSH HOLLA B. received the B.E. degree from the Srinivas Institute of Technology, VTU, Belgaum, and the master's degree in computer science and engineering from NMAMIT, India. He is currently an Assistant Professor with the Department of Data Science and Computer Applications, MIT, MAHE. His areas of interest are object detection, re-identification, and deep learning for computer vision.

MANOHARA M. M. PAI (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering. He is currently a Senior Professor with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. He has a rich experience of 33 years in Teaching/Research. He holds ten patents to his credit and has published 160 papers in National and International journals/conference proceedings. He is also a principal investigator for multiple industry/government research projects. He has published two books, guided six Ph.D.'s, and 85 master's thesis. His research interests include data analytics, cloud computing, the IoT, computer networks, mobile computing, scalable video coding, and robot motion planning. He has received the National Technical Teacher Award from the Ministry of Education, Government of India in 2022. He has been an Executive Committee Member of the IEEE Bangalore Section, Mangalore Subsection, and the Past Chair of the IEEE Mangalore Subsection. He is also a Life Member of ISTE and a Life Member of the Systems Society of India.

UJJWAL VERMA (Senior Member, IEEE) received the M.S. (Research) degree in signal and image processing from IMT Atlantique, France, and the Ph.D. degree in image analysis from the Télécom ParisTech, University of Paris-Saclay, Paris, France. He is currently an Associate Professor with the Department of Electronics and Communication Engineering, Manipal Institute of Technology, India. His research interests include computer vision and machine learning for Earth observation images, with a focus on self-supervised learning, few-shot learning, semantic segmentation, and classification of UAV and satellite images and video. He is a recipient of the "ISCA Young Scientist Award 2017–2018" by the Indian Science Congress Association (ISCA), a professional body under the Department of Science and Technology, Government of India. He is the Co-Lead for the Working Group on Machine/Deep Learning for Image Analysis of the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. He is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He is also a Guest Editor for Special Issues in IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, and a Reviewer for several journals (IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS). He is also a Sectional Recorder for the ICT Section of the Indian Science Congress Association from 2020 to 2024. He is a Life Member of the Indian Science Congress Association.

RADHIKA M. PAI (Senior Member, IEEE) received the Ph.D. degree from the National Institute of Technology Karnataka, Surathkal, India. She is currently a Professor and the Head of the Department of Data Science and Computer Applications, Manipal Academy of Higher Education, Manipal, India. She has a teaching and research experience of over 32 years. She has published 105 papers in national/international journals/conferences and has guided three Ph.D.'s and several master's thesis. Her research interests include data mining, big data analytics, character recognition, sensor networks, and e-learning. She was a recipient of the National Doctoral Fellowship from AICTE, Government of India. She was an Executive Committee Member of the IEEE Mangalore Subsection.