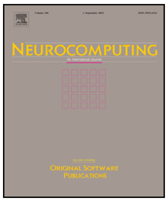
Neurocomputing 582 (2024) 127514

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



MSFFT: Multi-Scale Feature Fusion Transformer for cross platform vehicle re-identification

Ashutosh Holla B. [a], Manohara Pai M.M. [a,*], Ujjwal Verma [b], Radhika M. Pai [c]

[a] *Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India*

[b] *Department of Electronics and Communication Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India*

[c] *Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India*

ARTICLE INFO

Communicated by J. Lei

Keywords:
Surveillance
Vehicle re-identification
Self-supervised learning
Transformer
Keyframes

ABSTRACT

A vital component of Intelligent Transportation Systems (ITS) is vehicle re-identification, which allows vehicles to be identified across surveillance devices. Re-identification of vehicles is usually done using information collected from standalone surveillance devices such as fixed surveillance cameras (CCTVs) or aerial devices (UAVs). Re-identifying vehicles across standalone surveillance systems is challenging when there is a severe illumination change, a change of viewpoint, or an occlusion. Cross platform surveillance (CCTV+UAV) based vehicle re-identification is yet to be explored and can mitigate some of the challenges faced during re-identifying vehicles with standalone surveillance systems. This paper proposes a novel cross platform vehicle identification dataset called MCU-VReID using 42 CCTVs and a UAV. A novel re-identification method called Multi-Scale Feature Fusion Transformer (MSFFT) is proposed to re-identify vehicles observed across the cross platform surveillance systems. The network consists of inception layers with transformer networks that enable it to learn the vehicle's features at a variety of scales. The vehicles observed by two contrasting surveillance systems appear to be transformed representations of one another. Hence a two-stage training approach is facilitated for re-identifying vehicles observed across cross platform surveillance systems. The two-stage training approach aims to learn vehicle semantic transformations in the first stage using self-supervised approaches. The knowledge gained at the first stage relating to vehicle semantic transformations is transferred at the second stage of training to perform re-identification. Extensive experiments using the method demonstrate that MSFFT significantly improves over state-of-the-art methods to perform cross platform vehicle re-identification.

1. Introduction

In recent years video surveillance has emerged as an important aspect of providing robust security in traffic management. As a "smart-cities" initiative, surveillance devices are essential components of Intelligent Transportation Systems (ITS) that provide rich information required for robust traffic monitoring. Surveillance systems, mainly CCTVs are widely used for traffic monitoring. Surveillance footage contains a wealth of rich data that is beneficial for performing vision-related tasks such as the detection/counting of vehicles/pedestrians [1–3], re-identification and tracking vehicles/pedestrians [4–7], anomaly detection [8–10], etc. However, as an application of ITS, vehicle re-identification has attracted various researchers in computer vision. Vehicle re-identification aims to obtain a possible match of a vehicle observed in one camera with images of the exact vehicle appearing at different non-overlapping cameras [11]. In comparison with person re-identification, vehicle re-identification poses several challenges, namely: (a) there are limited appearance cues apart from vehicle color to differentiate vehicles of the same type/model, (b) Identical vehicles appearing at multiple cameras are subject to different viewpoints which lead to falsely re-identifying the vehicle of interest.

Usually, cameras are geographically distributed over an area under surveillance to perform vehicle re-identification. A vehicle observed at a specific CCTV is queried across remaining CCTVs to determine its existence, thereby drawing the trajectory of vehicle movement in

[*] Corresponding author.
E-mail addresses: ashutosh.b@manipal.edu (Ashutosh Holla B.), mmm.pai@manipal.edu (Manohara Pai M.M.), ujjwal.verma@manipal.edu (U. Verma), radhika.pai@manipal.edu (R.M. Pai).

https://doi.org/10.1016/j.neucom.2024.127514
Received 6 April 2023; Received in revised form 13 January 2024; Accepted 4 March 2024
Available online 12 March 2024

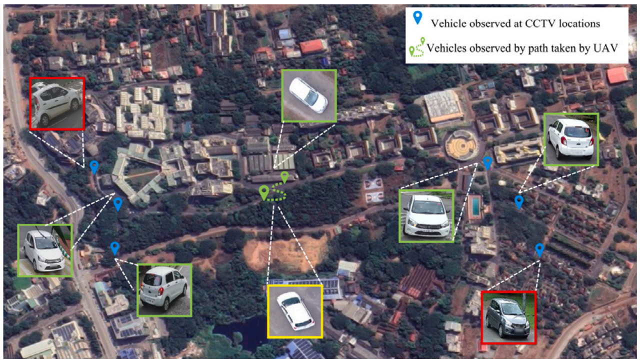

Vehicle observed at CCTV locations
Vehicles observed by path taken by UAV

Fig. 1. The dataset is acquired using 42 surveillance cameras and a UAV from an educational institution (188 acres). By utilizing a network of cameras and a UAV, vehicles are observed from multiple viewpoints and scales. cross platform vehicle re-identification aims at re-identifying a vehicle that is queried (vehicle image with yellow border) across a network of CCTVs and UAV. With significant viewpoint and appearance changes, cross platform vehicle re-identification retrieves a similar set of vehicles (vehicle image with green border) and a dissimilar set of vehicles (vehicle image with red border).

the event of vehicle tracking. Re-identification models employ appearance features, such as color, pose, and additional vehicle information, to identify identical vehicles captured by other surveillance cameras. However, the re-identification of vehicles done using CCTVs poses additional challenges since surveillance cameras may be vandalized during a public event or even be deactivated for maintenance purposes. Setting up CCTVs to monitor vehicle movements in the event that spans a short duration is expensive due to system expenses, extensive cabling, and labor charges. Under these circumstances, to minimize the cost, a handful of CCTV cameras are installed in selected traffic areas, creating a "blind spot". With this, there is a possibility that vehicles might be observed from the same viewpoint leading to less information about a vehicle for re-identification across other surveillance cameras.

Due to the rich characteristics of unmanned aerial vehicles (UAVs), their application for traffic monitoring has gained popularity among various countries [12]. UAVs are installed with cameras that allow them to monitor traffic movements with broader coverage and higher resolutions than static CCTVs. UAVs can be used to re-identify vehicles, presenting a wider range of perspectives. Because of their mobility, they can collect rich contextual information about vehicles. In addition, UAVs as a surveillance system exhibit particular challenges, such as restricted flights during night hours or in unpredictable weather conditions, limited battery backup, and high maintenance costs. Additionally, the change in flight altitude allows the vehicles to be observed and captured at a different scale which may lead to inaccurate re-identification of vehicles.

Recently re-identification of entities is performed using different image forms namely RGB, IR, NIR (Near Infrared) and thermal Infrared (TIR) acquired by different image acquisition surveillance cameras. These works [13–15] that uses these complementary image forms of an entity term the re-identification problem as "multi-modal" re-identification. The intuition behind these approaches is to addresses the limitation when handling single modality RGB entity images that are sensitive to illumination and environmental factors. The present study mainly focuses on how efficiently the re-identification of vehicles can be performed using the data acquired by two different platforms, viz. CCTVs and UAV. The vehicle appearing in each of the surveillance systems appears to have a significantly different viewpoint and appearance in these contrasting views. In this work, a novel cross platform vehicle re-identification framework is outlined to re-identify vehicles acquired by CCTVs and UAV. To the extent of our knowledge, vehicle re-identification using two different surveillance systems together is unexplored, and there is no publicly available vehicle re-identification dataset. A novel dataset called the Manipal CCTV-UAV Vehicle Re-Identification Dataset (MCU-VReID) is been developed to facilitate cross platform vehicle re-identification. It is obtained from an educational institution (Fig. 1) and contains 51 identical vehicles observed by a network of surveillance cameras and a UAV. For 51 identical vehicles, a total of 5630 manual annotations of vehicles are performed.

A novel re-identification method named Multi-Scale Feature Fusion Transformer (MSFFT) is proposed to re-identify vehicles across a network of CCTVs and UAV. MSFFT consists of several levels of inception, with transformer blocks that are used to extract vehicle features at different scales. Each level of MSFFT has inception layers of various filter sizes that extract features that are then transmitted to the transformer network at the next level. The transformer network further enhances the generalization of vehicles by computing attention scores for different scales at each level of MSFFT. Through transformer networks, more discriminative and scale-invariant features of vehicles can be learned for better re-identification. The vehicle features learned at different scales are fused to perform the re-identification of vehicles seen across CCTVs and UAV. A vehicle observed by two contrasting surveillance systems appears to be the transformed/augmented form of the other. Using the traditional supervised method of learning these transformations will not guarantee better re-identification. Thus, the proposed method employs a two-stage training strategy (Fig. 2) where the algorithm learns vehicle transformations in a self-supervised approach and utilizes these representations to perform re-identification as a downstream task. Most self-supervised methods use heavy data augmentation during the training phase so that the model is invariant to any transformation. This study avoids expensive data augmentation by considering vehicle images observed by either surveillance platform as the augmented version of the other (vice versa). The learned representation is then used to perform re-identification as a downstream task. The main contribution of the study is summarized as follows:

- A novel dataset to perform cross platform vehicle re-identification by acquiring CCTVs and UAV aerial videos at a gated university campus.
- A vehicle re-identification framework, namely a Multi-Scale Feature Fusion Transformer (MSFFT) is developed to re-identify vehicles across cross platform surveillance systems. MSFFT for cross platform vehicle re-identification learns the multi-scale features of vehicles at different levels by **integrating** the feature map produced by inception layers with transformer encoder layers. The integrated feature extractor network with inception layers and transformer encoders enhances vehicle re-identification by capturing multi-scale features and calculating semantic attention to vehicle parts, enabling a more enhanced and adaptable understanding of distinctive characteristics of vehicles required for cross platform vehicle re-identification.
- A two-stage training strategy to re-identify vehicles observed in cross platform surveillance systems. At first stage, the network is exposed to learn the transformation of vehicles observed by CCTVs and UAV using self-supervised approaches. The images of vehicles observed by UAV/CCTVs are treated as augmented views of other vehicles, thus avoiding the costly data augmentation required for self-supervised approaches. Using supervised learning, the knowledge gained at the first stage of training is then transferred to perform re-identification at the second stage.

2. Related work

This section presents recent contributions to vehicle re-identification performed with both CCTV and UAV surveillance systems. A detailed overview of the publicly available vehicle re-identification dataset is provided in Section 2.1. Further, Section 2.2 provides a summary of various contributions made to perform re-identification using either CCTVs or UAVs. As the present study uses a self-supervised approach to learn the transformation of vehicles, Section 2.3 summarizes the recent self-supervised techniques applied in vehicle re-identification.

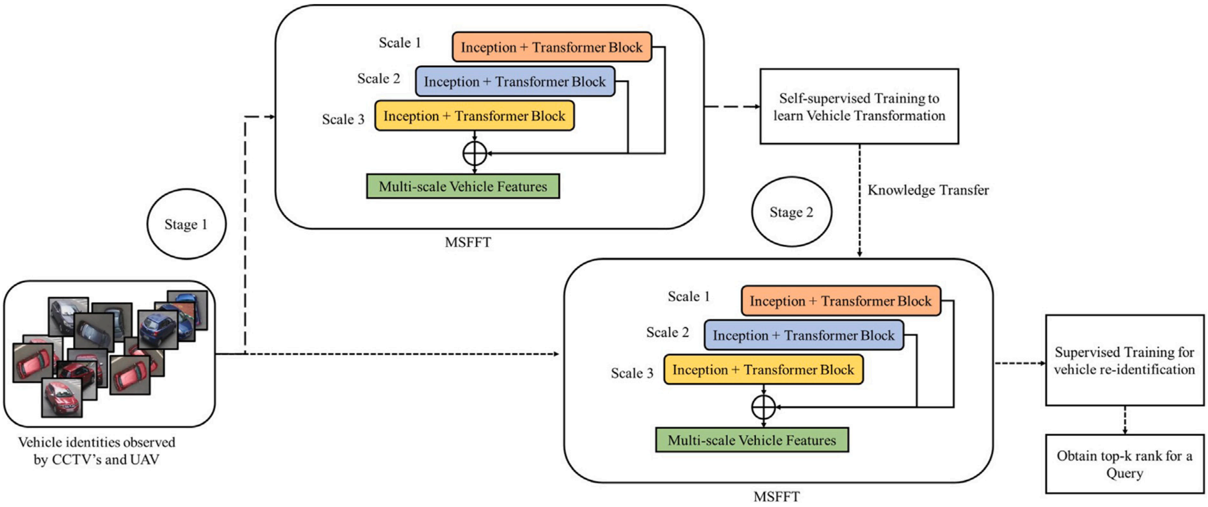

Scale 1
Inception + Transformer Block
Scale 2
Inception + Transformer Block
Scale 3
Inception + Transformer Block
Multi-scale Vehicle Features
MSFFT
Self-supervised Training to learn Vehicle Transformation
Knowledge Transfer
Stage 1
Stage 2
Vehicle identities observed by CCTV's and UAV
Scale 1
Inception + Transformer Block
Scale 2
Inception + Transformer Block
Scale 3
Inception + Transformer Block
Multi-scale Vehicle Features
MSFFT
Supervised Training for vehicle re-identification
Obtain top-k rank for a Query

Fig. 2. A two-stage training strategy for re-identifying vehicles observed by CCTVs and UAV. At the first stage of training, the MSFFT is exposed to learning the vehicle transformations using self-supervised approaches. At the second stage, the knowledge gained by learning various transformations at the first stage is transferred to train the MSFFT in a supervised approach for re-identification.

2.1. Datasets

2.1.1. CCTV vehicle re-identification datasets

CompCars

Yang et al. [16] contributed a large-scale vehicle dataset named CompCars. This comprehensive vehicle dataset comprises 2,14,345 images of 1716 car models. The vehicle images are collected from the web and surveillance cameras. For vehicles observed at different viewpoints, the dataset is labeled with five viewpoints that include front, rear, front-side, and rear side. The vehicle images are annotated with a bounding box and additional information, such as the model and color of the car. The dataset is applicable for car classification and verification and attributes prediction.

VehicleID

The vehicleID [17,18] dataset consists of 2,21,763 images in total, with 26,267 vehicles observed across multiple nonoverlapping surveillance cameras. Vehicle images are annotated with the help of the vehicle's license plate number. In the training set, the dataset comprises 1,00,182 images of 13,164 vehicles, and a test set consists of 20,038 images of 2400 vehicles. These vehicle images are either captured from the rear or from the front.

VeRi-776

Liu et al. [19] developed VeRi-776 which consists of 51,035 images of 776 vehicles observed across 20 surveillance cameras. The training set, consists of a total of 576 vehicles with 37,778 vehicle images. For the test set, the dataset comprises 200 vehicles with 11,597 vehicle images. 1678 vehicle images from the test set are used as query images. The vehicles are labeled by providing a bounding box over the entire vehicle, along with type, color, and vehicle correlation.

VERI-Wild

Lou et al. [20] contributed to a large-scale vehicle re-identification dataset acquired by 174 surveillance cameras over a month in unconstrained scenarios. The dataset consists of 4,16,314 vehicle images of 40,671 vehicle identities. The bounding box is generated using YOLO-v2. The dataset is split into training and testing, which comprises 30,671 vehicles with 2,77,797 images for training and three subsets for testing.

2.1.2. UAV vehicle re-identification datasets

Vehicle re-identification in aerial videos is under explored, and there is a limited standard dataset for vehicle re-identification from videos acquired from UAVs. This section briefly discusses UAV videos based on vehicle re-identification datasets.

VRAI

In recent years, a large-scale dataset has been developed specifically to perform vehicle re-identification using the data acquired by aerial devices [21]. The dataset consists of 137K images of 13K vehicle instances observed by two UAV surveillance devices. The UAVs' flight altitude ranges from 15 m to 80 m. Manual annotation has been provided, which includes vehicle type, color, and different vehicle parts. A total of 350 pairs of video clips are obtained with a total duration of 34 h.

UAV-VeID

A vehicle re-identification dataset acquired by UAV surveillance devices has been contributed in [12]. The dataset was acquired using two UAVs comprising 4601 vehicles, 41,967 annotated vehicle images, and 16,850 query images of vehicles. Each of the vehicle images is taken from different viewpoints. In addition, the dataset contains a distractor set composed of 300K falsely detected bounding boxes and vehicles that do not belong to 4601 annotated vehicle identities.

The Table 1 summarizes the datasets discussed above to perform vehicle re-identification are acquired by either standalone surveillance devices, i.e., CCTV or UAVs. These datasets contain vehicles subjected to illumination changes, variance in scale, viewpoint changes, etc. In addition, some of these datasets provide supplementary information such as the type/model of the vehicle, its color, or its pose, which may aid in generating a re-identification framework for better re-identification. In exchange for this, additional labor is required to provide this information through data annotations and labeling.

With cross platform surveillance cameras, vehicles are captured at dynamic scales and resolutions from multiple viewpoints, enabling them to gather both aerial (UAV) and ground-level views (CCTVs). Consequently, more information about vehicles is collected, which allows them to be re-identified more precisely. Moreover, cross platform surveillance systems can also be used to enhance re-identification when vehicles observed by either surveillance platform are exposed to viewpoint/illumination changes and partial occlusion. Since there is limited work contributing to perform vehicle re-identification through cross platform surveillance systems, a vehicle re-identification dataset is developed to facilitate the re-identification of vehicles observed across surveillance cameras and a UAV. In addition to the presence of challenges that occur with datasets specific to standalone surveillance systems, MCU-VReID introduces additional challenges: a significant

Table 1
Comparison of the existing dataset with the proposed dataset for vehicle re-identification.

Dataset	Year	Mode of data gathering	# cameras (CCTVs/UAV)	# Vehicle identities	Number of images	Supplementary information
CompCars	2015	Web + Surveillance camera	–	1716	2,14,34	Yes
VehicleID	2016	CCTVs	Multiple cameras	26,267	2,21,763	Yes
VeRi-776	2016	CCTVs	20 CCTVs	776	51,035	Yes
VERI-Wild	2019	CCTVs	174 CCTVs	40,671	4,16,314	Yes
VRAI	2019	UAVs	2 UAVs	13,000	~137K	Yes
UAV-VeID	2021	UAVs	2 UAVs	4601	41,917	Yes
MCU-VReID	Proposed	**CCTVs and UAV**	**42 CCTVs + UAV**	51	5630	**No**

transformation of vehicle images taken from two contrasting views, limited overlapping vehicle parts for images acquired by CCTV's and UAV, etc. The dataset is an extension of the work [22] that was conducted to perform cross platform vehicle re-identification. The developed dataset avoids providing supplement information such as license plate information, vehicle type/model, pose and color information making it more challenging to design a framework for re-identification. Further details of the MCU-VReID dataset are described in Section 3.1.

2.2. Existing work on re-identification

2.2.1. Person re-identification

The study presented by authors in [23] generates a convolutional descriptor of part features from a Part Based Convolution (PCB) network. The framework utilizes the standard ResNet-50 backbone architecture to learn the vehicle representations. PCB splits the feature maps horizontally into 'p' parts to take advantage of local spatial cues. Each part is later fed as an input to the classifier to predict the person's identity.

In the work [24], the authors designed a CNN framework termed an Omni-scale network that consists of residual blocks of several convolutional streams for detecting the features at a different scale. The authors used depth-wise separable convolutions, which initially perform a pointwise convolutional operation followed by a depth-wise convolution. They introduced a unified aggregation gate that dynamically fuses multi-scale features with channel-wise weights.

The authors of the work in [25] presented a novel approach called inter-instance contrastive encoding (ICE) for unsupervised re-identification based on contrastive learning. ICE addresses enhancing the quality of pseudo-label generation used in contrastive learning by leveraging inter-instance entity similarity scores. ICE uses two components namely hard and soft instance contrastive loss to minimize intra-class re-identification variance and to enhance the network to be more consistent with augmented and original views of identity.

A Part-based Pseudo Label Refinement (PPLR) is developed by authors in [26] and aims to reduce label noise by determining complementary relationships between global and local features. The framework uses k-nearest neighbors to measure similarities between global and local part features. The framework aims to learn the discriminative features of vehicles when addressing datasets that have minimal label information about entities.

In the study [27] authors developed a re-identification framework named Neighbor Transformer (NFormer) that explicitly models the interactions between all the input images for learning the representations. It includes two modules namely Landmark agent attention (LAA) and Reciprocal Neighbor Softmax (RNS) to determine a relationship map of images in a feature space. The framework uses these modules to eliminate outliers in the feature space for robust re-identification.

Authors in [28] proposed a novel Part-aware Transformer (PAT) model for domain generalization and person re-identification (DG-ReID). The PAT framework is developed to learn generic features that are robust to domain-specific representations. The framework uses two key components namely Cross-ID similarity learning (CSL) and Part-guided Self-Distillation (PSD). CSL of the PAT framework learns to label invariant features by mining local visual information shared by different labels. This knowledge of local visual representation is utilized by the PAT to guide in learning the global features.

2.2.2. CCTV vehicle re-identification

By utilizing publicly available datasets, several works have contributed to vehicle re-identification. To aid in the re-identification process, some of the methods utilize vehicle attributes or additional information. This information are either directly obtained from the dataset or labeled for task-specific purposes. Some of the works that incorporate this supplement information for re-identification are summarized below.

The authors in the work [17] proposed a two-branch deep convolutional network that projects the vehicle images to a euclidean space to measure the similarity of the two vehicles. To learn the discriminative features from deep feature embeddings, the network utilizes a Coupled-Cluster Loss (CCL).

To alleviate the requirement of labeled data, the authors in [29] proposed an adaptive feature learning method to perform re-identification. A re-identification network is trained on existing datasets by fine-tuning the feature extractor module to adapt any different target dataset. Their proposed framework consists of three stages: the vehicle proposal stage, single-camera tracking, and a feature extractor step to perform re-identification.

Utilizing the vehicle part information, the authors in [30] proposed a re-identification framework that exploits the challenges of distinguishing vehicles with similar geometric shapes and appearances. To address the near-duplicate phenomenon, local and non-local features must jointly learn the vehicle features. Their framework incorporates a local module that gives importance to vehicle parts. YOLO object detector is used to detect and locate the vehicle parts. Later these vehicle parts are projected into a global module that utilizes a ResNet-50 [31] as a backbone network to learn the feature embeddings.

The success of transformers in computer vision applications led authors in [32] to propose a transformer-based object identification system named TransReID. Their approach to performing re-identification includes two modules: a jigsaw patch module (JPM) and a side information embedding module (SIE). The jigsaw patch module shuffles the generated patch embeddings to capture robust object features when it is subjected to partial occlusion. The SIE also learns non-visual cues by introducing additional embedding representations such as camera or viewpoint information. Authors evaluated their approach using different variants of DeiT [33] and ViT [34] transformers on benchmark person and vehicle re-identification datasets.

A novel loss paradigm named Sparse Pairwise loss (SP) is developed by authors in [35] to address the drawbacks of pairwise loss for object re-identification. The SP loss constructs a sparse similarity matrix for each class to consider a few appropriate image pairs in a mini-batch. Further, the SP loss mines only top-k negative and positive pairs to ensure that only informative pairs are chosen for training. SP loss also uses an adaptive mining strategy to adapt to large intra-class instance variations.

Apart from the above summarized works, several re-identification methods are presented that aim to perform re-identification with minimal use of supplementary information about vehicles.

An unsupervised metric learning model is developed that leverages pairwise and triplet constraints on training a re-identification model using the triplet loss similarity metric [36]. The vehicle features are transformed from an initial input dimension into a feature space where similar identity vehicles are close together while keeping dissimilar

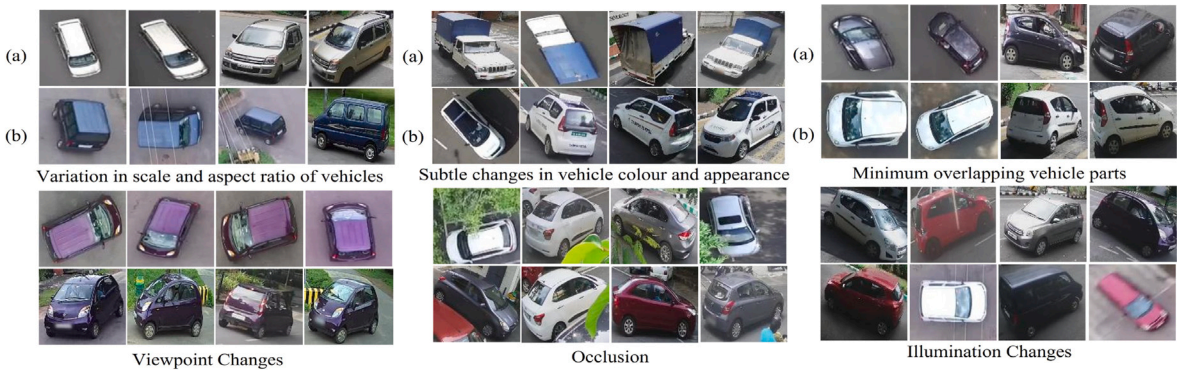

(a)
(b)
Variation in scale and aspect ratio of vehicles
(a)
(b)
Subtle changes in vehicle colour and appearance
(a)
(b)
Minimum overlapping vehicle parts
Viewpoint Changes
Occlusion
Illumination Changes

Fig. 3. Illustration of different challenges presented in MCU-VReID. Top: Each row (a) and (b) shows the vehicle identities observed by the network of cameras and UAV. Re-identification is more difficult due to challenges such as differences in aspect ratio, scale, appearance, and minimum overlap of vehicles when observed by CCTVs and UAV. Bottom: MCU-VReID also presents scenarios of an identical vehicle subjected to viewpoint changes, occlusion and illumination changes.

vehicle identities far apart. A Single-shot detector [37] is utilized to identify the vehicles appearing in a scene and is assigned to an existing or a new tracklet. The detector is built upon a VGG-16 [38] backbone network and is trained using a COCO dataset with only vehicle class. To compare the similarity between two vehicles, a middle frame of the tracklets is selected, and the similarity score is computed using Euclidian distance.

The authors in [39] introduced a batch sampling strategy with triplet loss to perform vehicle re-identification. They evaluated the batch sample and batch weighted variants against the standard batch hard and batch all variants of vehicle re-identification.

A vehicle re-identification and abnormality detection framework were contributed in work [40]. The framework consists of three steps: a deep metric embedding module, a vehicle classifier module, and a re-ranking module. The deep metric embedding module is utilized to extract discriminative vehicle features. The classifier module addresses how the vehicle features can be learned when they are of different poses and colors. A Faster-RCNN is used to detect the vehicles appearing on the scene. The features of detected vehicles are learned using a ResNet-50 [31] trained with a triplet loss metric. As a post-optimization step, the authors have re-ranked the candidate images for a given query using the bag-of-words approach.

Apart from these works, some recent studies considered viewpoint cues to enhance the conduction of vehicle re-identification [41–44]. As a part of the AI CITY Challenge [45,46], several other vehicle re-identification works have been contributed [47–50].

Though the above approaches aim to address different challenging aspects of vehicle re-identification, there are some hurdles that may arise while performing cross platform vehicle re-identification. Some of the above approaches consider deep feature extraction networks with fixed filter sizes. Due to that, the feature extractor uses only local neighbor information at a time. Further, deep feature extractor networks suffer information loss when they contain more strided convolution and downsampling operations. On the other hand, a transformer-based re-identification approach learns vehicle features by computing attention scores at a fixed patch size. Further existing approach involves additional computation in determining camera/viewpoint embeddings for learning the features of vehicles observed from different viewpoints.

2.2.3. UAV vehicle re-identification

There are limited works carried out in performing vehicle re-identification using aerial data. The authors in [21] proposed a method to perform re-identification that consists of two parts: First, a Multi-objective model whose backbone network is ResNet-50 is used for retrieval, ID classification, and attribute classification. Secondly, discriminative parts of vehicles are identified using YOLOv2 [51].

A vehicle re-identification dataset using UAV named UAV-VeID is developed by [12]. To address substantial changes in viewpoints and scales of vehicles, authors in [12] have proposed a viewpoint adversarial training strategy and a multi-scale consensus loss to promote the robustness and discriminative power of learned deep features. They have used dilated convolution concept at different dilation rates to address the problem of variation in the scale of vehicles when observed by UAV at different altitudes levels. Research towards vehicle re-identification using UAV is emerging but significantly less than re-identification using surveillance cameras. Vehicles observed by UAV's differ in scale as the flight altitude of UAV's may vary dynamically. Existing vehicle re-identification approaches tackle to address the scale-invariant for better re-identification, but they fail to address how re-identification can be performed when there is no significant overlap of vehicle information or contrasting appearance of vehicles observed by two different surveillance systems.

2.3. Self-supervised based training for vehicle re-identification

Self-supervised methods are emerging to address scenarios where collecting and annotating large-scale datasets is expensive for performing computer vision tasks. Self-supervised methods utilize a contrastive learning approach that aims to cluster similar examples together and distancing from diverse samples. Traditional self-supervised methods define a pretext task for underlying deep learning models to learn visual features from unlabeled data [52]. After the model has been trained to identify visual features, the features learned from the initial layers of the model are transferred to train downstream tasks. A typical autoencoder architecture is used to solve the pretext problem. In recent years several works related to self-supervised learning have been presented by various authors. Some of the methods, such as SimCLR [53], MoCo [54], BYOL [55], SwAV [56] use different techniques to train the network namely data augmentation, memory banks, momentum encoders, clustering, etc.

In the context of vehicle re-identification, an approach to learn vehicle-specific discriminative features called Self-supervised Attention for Vehicle Re-identification (SAVER) is proposed by authors in [57]. A Variational Autoencoders-based reconstruction module generates the vehicle image without specifically identifying the vehicle. The reconstructed image is further masked from the input image to produce an image that contains vehicle-specific information. Additionally, this is provided to a feature extraction module for training re-identification with triplet and cross-entropy loss.

To learn geometric features of vehicles in a self-supervised approach, authors in [58] presented a framework that encodes geometric local features and global contextual information. They have used a self-supervised approach to learn the geometric features of vehicles from different perspectives. The framework comprises of the global branch, a self-supervised learning branch, and geometric features branch to encode local and global geometric and discriminative features of vehicles. Authors evaluated their framework with a publicly available re-identification dataset namely VeRi-776 [19], VehicleID [17] and CityFlow-ReID [59].

Authors in [60] presented a re-identification framework called Self-Supervised and Boosted Vehicle re-identification. The framework is based on the principle of student–teacher learning with a momentum encoder. The student network is modeled to learn the local views of the vehicle, while the teacher network is trained to learn the global view. The student network is optimized with self-supervised (sharpening and centering) and re-identification strategies (Triplet Loss).

Recently several authors have also explored the problem of cross domain vehicle re-identification. cross domain vehicle re-identification addresses the challenge of adapting re-identification methods to a target domain that has a drastically different scene, different vehicle types/models, variation in viewpoints, illumination changes, etc. which has no commonalities with the source domain for which the framework was designed. Cross domain approach for vehicle re-identification bridges this gap of re-identifying the vehicles of a target domain by domain adaptation strategies such as GANs for target domain adaptation [61,62], attention gates for learning more discriminative features of target domain using progressive feature learning [63] and also the use of unsupervised learning through CycleGANs [64]. The performance of these cross-domain frameworks is assessed using existing datasets that exhibit significant variations in scene and vehicle characteristics but are obtained using a single surveillance system. In contrast, the present study addresses cross platform vehicle re-identification of vehicles that are observed by two different platforms of surveillance systems i.e. CCTVs and UAV. In the present study, the re-identification model utilizes data that includes vehicles viewed at the common set of scenes but gathered by two different surveillance systems, as opposed to the cross domain approach. Though the vehicles are seen in a similar set of scenes, the scene information is taken by two *distinct platforms of surveillance systems* that vary significantly specific to the resolution, appearance, and viewpoints of vehicles.

For vehicle re-identification, most approaches developed are concerned with re-identifying vehicles observed by CCTV or UAVs. In the re-identification of vehicles, supervised training is typically used with/without supplementary information. However, in the present study since the vehicles are observed by two contrasting surveillance systems, there is a drastic change in the semantics of the vehicle. Re-identification frameworks trained in a supervised approach for vehicle images will fail to generalize the features of vehicles observed by UAVs and CCTVs. Some contributions have been made to the re-identification of vehicles observed by cross platform surveillance systems. In [22], an attempt is made to re-identify vehicles appearing in two different modes using a supervised approach. Here, re-identification is conducted by explicitly computing transformation using homography between vehicle images acquired by CCTV and UAV. Subsequently, the transformation is applied to CCTV vehicle images to obtain a near-matching vehicle image if a UAV takes it. However, re-identification appears to be challenging due to various factors. In certain cases, it is found that due to poor estimation of corresponding points, the computed homography is not accurately calculated to perform the required transformation of vehicle images for re-identifying vehicles. Also, vehicles observed from different perspectives made it more challenging for CNN to learn vehicle features. This minimized the possibility of obtaining cross platform candidate images for a given query. Consequently, in the present study, the distribution of vehicle features observed by CCTVs and UAV is assumed to be different. By learning the appropriate transformations using self-supervised approaches, the similarities between vehicle images observed by UAVs and CCTVs are minimized. The knowledge gained through a self-supervised approach is then used for supervised re-identification. The details of the framework and training strategy are presented in Section 3.

3. Proposed work

The following sections present the details of the contributions of the article mainly providing detailed information on the developed dataset i.e. MCU-VReID (Section 3.1) and vehicle re-identification framework MSFFT (Section 3.2) for performing cross platform vehicle re-identification.

3.1. MCU-VReID

Existing vehicle re-identification methods are evaluated on datasets developed using standalone surveillance systems i.e., CCTV or UAV. Building upon the work conducted in [22], this study further explores the task of vehicle re-identification using cross platform surveillance systems. For this, a dataset called Manipal CCTV-UAV Vehicle Re-Identification (MCU-VReID) is developed. The primary significance of the dataset is to perform cross platform vehicle re-identification. The developed dataset is also helpful in performing other vision-related tasks such as vehicle detection, tracking, and identifying abnormal behaviors.

The surveillance data is acquired from an educational Institution (Manipal Institute of Technology, Manipal, India). From the entire available CCTV, a total of 42 CCTV cameras were chosen. A fleet of cameras is considered from the Academic area of the campus, and other cameras are located at hostel premises thereby covering the entire campus. It may be noted that traffic is not uniform in all the cameras. In an earlier work [22] for cross platform vehicle re-identification, data was collected from 20 CCTVs and a UAV on campus, mainly from the Academic area. In this study, the dataset is further extended to cover a larger geographical area: along with 20 CCTVs located in the academic area, additional 22 CCTVs from the hostel area of the campus are considered. The videos acquired from UAV and CCTVs at the hostel premises present significant challenges specific to illumination changes. Using the chosen CCTVs and UAV, the data is acquired on different days. UAV aerial videos are acquired using a DJI Phantom Professional drone with a 1280 × 720 resolution at 29 frames per second. While acquiring the videos using UAV, the flight altitude is varied between 25–30 m. The CCTV videos are captured with 1920 × 1080 resolution at 20 frames per second. The videos have been acquired such that the vehicles a UAV observes on a flight are seen by at least one or two CCTV cameras. During data acquisition, it is made sure that when a UAV is made to fly across the locations of installed CCTV cameras for acquiring vehicle information for a certain duration, the data from CCTV videos are acquired at the same timestamp. The CCTV videos were also acquired in certain scenarios before the UAV flight and post-landing. This is to collect the vehicles in motion previously observed by UAVs in their respective parking lots. The collected UAV video duration ranges between 8 to 12 min, while the CCTV video duration ranges from 15–40 min. The dataset consists of 51 vehicles observed across CCTVs and a UAV. The summary of the dataset is shown in Table 2.

Most of the existing re-identification tasks use available datasets containing supplement information which eases the re-identification task. This supplement information can be the model/type of vehicle, the color of the vehicle, and pose information (front view, rear view, and side view) which provides a substantial cue of a vehicle observed by a camera. This supplement information requires additional labeling, which is expensive and time-consuming. Although the developed dataset using cross platform surveillance systems can help re-identify vehicles through aerial and ground-level views, it also presents certain additional challenges. Apart from common obstacles such as occlusion, illumination, and viewpoint changes, the dataset presents extra challenges when the vehicle is observed by both UAV and CCTVs. Among them are subtle differences in the color and appearance of certain vehicles, the minimal overlap of vehicle information, and differences in scale and significant transformations of vehicles as observed by CCTVs and UAV (Fig. 3). Based on the datasets developed for person re-identification that use minimal supplemental information, the dataset developed in this work follows a similar approach of providing minimal supplementary information to perform vehicle re-identification.

Table 2
Dataset information.

Surveillance platform	Surveillance cameras	Image resolution	Frame rate	Minimum duration (min)	Maximum duration (min)
CCTV	42	1920 × 1080	20	15	40
UAV	1	1280 × 720	29	8	12

3.1.1. Data pre-processing

In this study, vehicle re-identification involves querying a vehicle observed in either UAV or CCTVs. As described in Table 2, the surveillance videos representing both UAV and CCTV are taken at 29 and 20 frames per second, where the change between successive frames is marginal. As a pre-processing step, a shot boundary detection [65, 66] algorithm is applied to produce keyframes that contain identical vehicles observed across both platforms. The duration of surveillance videos varies from 8–12 min (UAV) and 15–40 min (CCTVs). The resolution of UAV and CCTV videos is 1280 × 720p and 1920 × 1080p respectively. Hence, generating keyframes using shot boundary detection with the original frame dimension is computationally expensive. To generate keyframes, every frame of surveillance videos is resized to a dimension of 512 × 512. In shot boundary detection, every frame is divided into a non-overlapping uniform grid. In order to identify a histogram difference between two corresponding frames with two consecutive windows, a Chi-square distance is used. The difference is calculated for every pair of consecutive frames, and a shot boundary is identified if the histogram difference between two frames is greater than a threshold T_{shot}. For every frame appearing in a shot of the shot boundary detection, a middle frame is designated as a keyframe. Later, the vehicle re-identification is carried out for the identified keyframes by considering it with the original image resolution of either platform ($1920 \times 1080p$ for CCTV and $1280 \times 720p$ for UAV).

For the generated keyframes, only those "identical vehicles" observed across both CCTVs and UAV are identified. These identical vehicles are numbered sequentially, resulting in a total of 51 identical vehicles. With an assumption that vehicles may undergo appearance changes (damaged vehicle part, re-paint, the addition of stickers, etc.), a similar vehicle observed on two separate days is assigned with a different vehicle identity. These vehicles are manually identified with the help of the vehicle's license plate information. Later the license plate details of every vehicle appearing in the keyframes are blurred using an image-blurring tool. The identical vehicles are individually labeled by drawing a bounding box and assigning a vehicle identity for re-identification. Providing annotation for identical vehicles is manually performed using the Microsoft Visual Object Tagging tool (MS Vott).

3.2. Methodology

3.2.1. Overview

The present work performs vehicle re-identification across cross platform surveillance systems. Vehicles observed by two contrasting surveillance systems (UAV and CCTVs) may exhibit several appearance changes and variations in scale, making it challenging to re-identify the vehicles. To address these, a novel vehicle re-identification framework is designed to re-identify vehicles observed by a network of CCTVs and a UAV. Fig. 4 shows the proposed Multi-Scale Feature Fusion Transformer (MSFFT) that comprises convolution layers and blocks of Inception layers to capture the contextual information of vehicles at different scales. The learned features at each Inception block are propagated to multi-head attention layers that compute attention among feature maps at different resolutions. The feature maps generated at different resolutions are fused to obtain a representation that accumulates the features at different scales. The fused representations are used to generate feature embeddings that are used further for two-stage training. During inference of vehicle re-identification, for any given query vehicle image, the features of the query vehicle and the features of all the vehicle images from the gallery are extracted. A similarity matrix is generated for each query to the gallery vehicle images based on a distance metric. For every query image, the gallery images are then ranked according to the similarity matrix. The similarity matrix are sorted such that the most similar vehicles having a higher rank than the rest of the vehicles with lower similarity to the given query.

In the present study, the vehicles observed by CCTVs and a UAV show drastic differences in appearance. Fig. 5 shows that a vehicle observed by a CCTVs is a transformed/augmented version of the same vehicle spotted by a UAV or vice versa. Fig. 5 shows that the vehicle observed by UAV and CCTVs contains some common overlapping parts information. For example (a) from Fig. 5, the vehicle contains similar overlapping rear portions of the vehicle observed by UAV and CCTV. The vehicles appearing in examples (b), (c), and (d) of Fig. 5 share a small amount of similar information from the front and side view of the vehicles. Therefore, the underlying re-identification network must be exposed to learn the different semantic perturbations. In doing so, the network can map two different contrasting views of vehicles image taken by a CCTVs and UAV aerial device. For this study, two state-of-the-art self-supervised learning strategies are considered. Specifically, SimpleSiamese (SimSiam) [67] and Nearest Neighbor Contrastive Learning (NNCLR) [68] are used to guide the underlying re-identification network invariant to any augmentations or transformations.

A two-stage training strategy is adopted to train the MSFFT. The identical vehicles observed by UAV and CCTVs have significant appearance changes. The vehicles observed by surveillance platform (CCTV/UAV) appear to be a transformed version of the other. Hence the model needs to be invariant to these transformations to accurately re-identify vehicles. The main objective of self-supervised approaches is to expose the underlying network to various data enhancements so that it is invariant to future vision tasks. Self-supervised approaches typically consist of an encoder and a predictor network (Fig. 6). Various data augmentations are applied to input batches, and a correlation is estimated between the original and the augmented view of the image. Using a contrastive loss, the distance between an encoded image feature embedding and an augmented view of the same image is smaller than the distance between the augmented views of different images. In contrast, this study assumes that the exact vehicle observed by either UAV or CCTV represents an augmented/transformed view of the other. Therefore the study avoids expensive data augmentations which is a laborious task and in some cases may even change the semantic information of the vehicle. The MSFFT framework is initially trained using two self-supervised approaches: SimSiam (Simple Siamese) and NNCLR (Nearest Neighbor Contrastive Learning). These two approaches are implemented independently. Later the framework is trained by fine-tuning the final layers to perform re-identification. The following subsections provide details regarding the architecture of MSFFT and the training strategy used to conduct re-identification.

3.2.2. MSFFT: Multi-scale feature fusion transformer

A novel vehicle re-identification framework named Multi-Scale Feature Fusion Transformer (MSFFT) is designed to re-identify vehicles across cross platform surveillance cameras. The framework has two modules. The first module namely multi-scale feature extraction, consists of inception layers [69] along with a transformer network to capture contextual information at different scales. The inception layers target learning the characteristics of vehicles of dynamic scales and sizes. This is done by applying the convolution operation to images

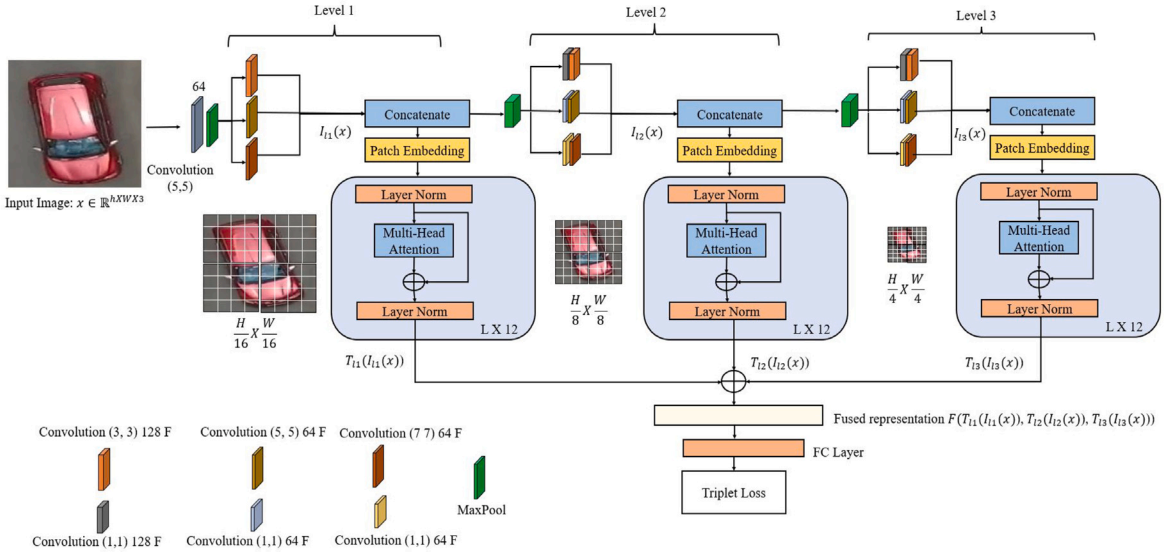


Fig. 4. MSFFT: Overview of the proposed architecture for cross platform vehicle re-identification. The architecture consists of blocks of inception layers with transformer networks that extract features at different scales. The vehicle features are learned by Inception layers with transformer networks at three levels which are fused to encapsulate the features of vehicles at multiple scales.

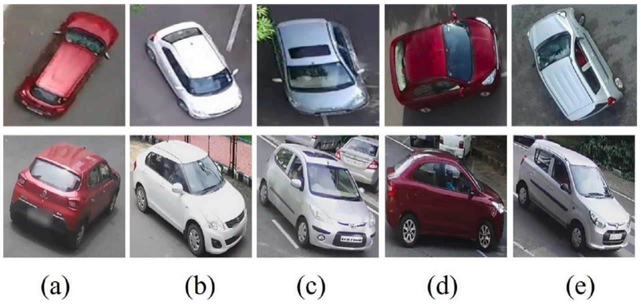


Fig. 5. Each column contains same vehicles observed by UAV (top tow) and CCTV (bottom row). Note that very few parts of the vehicles overlap in the two images. Thus, due to the overlap of certain vehicle parts, the vehicles observed by (CCTVs/UAVs) seem to be transformed/augmented versions of the vehicle seen by (UAVs/CCTVs).

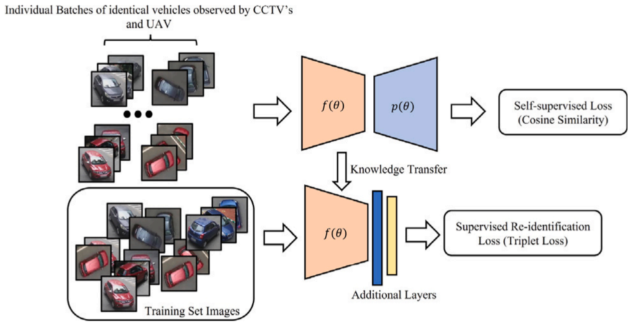

Individual Batches of identical vehicles observed by CCTV's and UAV
$f(\theta)$
$p(\theta)$
Self-supervised Loss (Cosine Similarity)
Knowledge Transfer
$f(\theta)$
Supervised Re-identification Loss (Triplet Loss)
Additional Layers
Training Set Images

Fig. 6. Two-stage re-identification approach comprises learning the vehicle semantics from an encoder ($f(\theta)$) and a predictor network ($p(\theta)$). The encoder ($f(\theta)$) is further used with additional layers to train a re-identification network in a supervised approach.

with different kernel sizes. The vehicle features learned by these inception layers are propagated to the transformer network for calculating attention scores. The transformer network calculates attention scores for a given resolution image patch. These inception layers, along with transformer networks, are used at three different levels (Fig. 4) to capture the contextual information of vehicles at multiple scales. The second module is a feature fusion module that fuses the features learned by the inception layers and transformer networks from different levels that contain the learned vehicle representations at multiple scales. The fused feature representation is further used as a feature embedding for training the network to learn the transformations using self-supervised approach and further use the learned representations to perform re-identification using supervised approach. The details of the MSFFT architecture are discussed below.

Multi-scale Feature Extraction:

MSFFT takes the bounding box of similar vehicles appearing across the keyframes of both CCTVs and UAV as input. Given a bounding box of vehicle $x \in \mathbb{R}^{h \times w \times 3}$, it is resized to an uniform dimension of 224 × 224 before feeding to the initial layers of the framework. The resized image x is passed to an initial 5 × 5 convolution layer with ReLU activation. The downsampled feature map is forwarded to the first level of multi scale feature extraction module comprising of inception like layers with transformer network. At the first level, the inception layers consist of three parallel convolution layers of various filter sizes, as shown in Fig. 4. The motivation behind using filters of various sizes is to capture vehicle features at various spatial locations in the image. Thus prominent features of vehicles of different scales and aspect ratios are learned using these inception layers. The feature maps from the first level of inception layers $I_{l1}(x)$ are concatenated and given as input to the transformer network. The concatenated feature maps of the inception layers $I_{l1}(x)$ have the dimension $H \times W \times F$ where H and W are the height and width of the feature map and F is the number of filters. The concatenated feature map from the Inception layers $I_{l1}(x)$ is further fed to the transformer block to compute attention scores with multi-head attention. The method uses the same operation as in standard Vision Transformer (ViT) [34] to compute multi-head attention over a sequence of patch embeddings. The patch embedding layer of ViT inputs a RGB image at the patch embedding layer. The patch embedding layer reshapes the image into a sequence of flattened 2D patches. In contrast, the proposed framework uses the output of each concatenated feature map of the inception layers to calculate attention scores. Hence the input to the patch embedding layer is the concatenated convolution feature maps of dimension $(H_{I_{l1}}, W_{I_{l1}}, F_{I_{l1}})$ where $H_{I_{l1}}$ and $W_{I_{l1}}$ are the height and width of the feature map of $I_{l1}(x)$ concatenated convolution feature map and $F_{I_{l1}}$ is the depth of the feature map. The patch embedding layer is similar to ViT, which divides the given input into $\frac{H_{I_{l1}}}{P} \times \frac{W_{I_{l1}}}{P}$ patches. The patch size for the first level of the transformer network is taken as $P = 16$. The generated patches are projected into a D dimension vector space to compute patch embeddings. Layer Norm (LN) is applied before every block and after the residual connections. Multi-head self attention is computed across $\frac{H_{I_{l1}}}{16} \times \frac{W_{I_{l1}}}{16}$ number of patches. The computed attention score by the transformer network at level 1 is denoted as $T_{l1}(I_{l1}(x))$.

The concatenated feature map of the inception layers $I_{l1}(x)$ from the first level of the multi-scale feature extraction module is downsampled using a maxpool operation with stride (2,2). This is further fed as input to the second level of the multi-scale feature extraction module. The inception layers at level 2 consist of three branches of two consecutive convolution layers. At each branch, the first convolution layer is a 1×1 convolution with filters of various sizes and depths, as shown in Fig. 4. Regular convolution layers follow it to calculate the contextual information for the downsampled feature map with different filter sizes (Fig. 4). The feature maps generated by three parallel convolution branches are again concatenated to generate the feature map $I_{l2}(x)$ for Inception layers at the second level of the multi-scale feature extractor module. The concatenated feature map is fed to the transformer network at the second level of the multi-scale feature extractor module to compute attention scores. At this level, the initial patch size is reduced by a factor of 2, thereby computing attention scores in a smaller patch neighborhood of $\frac{H_{I_{l2}}}{8} \times \frac{W_{I_{l2}}}{8}$. Here $H_{I_{l2}}$ and $W_{I_{l2}}$ denotes the height and width of feature maps from Inception layers $I_{l2}(x)$. The computed attention score by the transformer network at level 2 is denoted as $T_{l2}(I_{l2}(x))$.

At level 3 of the multi-scale feature extractor module the feature maps generated by the inception layers $I_{l2}(x)$ are downsampled in a similar way as performed in level 2. The downsampled features are fed to the final level of inception layers and the transformer network. The inception layers are similar to the inception layers at level 2. The features from the inception layers at level 3 $I_{l3}(x)$ are concatenated and fed to the transformer network to compute the attention scores at a patch of size $\frac{H_{I_{l3}}}{4} \times \frac{W_{I_{l3}}}{4}$. The computed attention score by the transformer network at level 3 is denoted as $T_{l3}(I_{l3}(x))$.

Feature Fusion Module:

The multi-scale feature extractor module of MSFFT aims at learning features at multiple scales. Vehicle features are learned using the inception layers and transformer network to compute attention scores at different feature map resolutions. The generated attention scores at three levels i.e. $T_{l1}(I_{l1}(x))$, $T_{l2}(I_{l2}(x))$, $T_{l3}(I_{l3}(x))$ contains contextual information about vehicles at different scales. These information are concatenated to generate a feature vector that encapsulates the vehicle features at multiple scales i.e.

$$Feature_{Fused} = concat(T_{l1}(I_{l1}(x)), T_{l2}(I_{l2}(x)), T_{l3}(I_{l3}(x))) \qquad (1)$$

where *concat* is the concatenation operation. In the two-stage approach to training the network, the fused feature vector generated as specified in Eq. (1) is forwarded to the layers that are used to train the network. The sub Section 3.2.3 presents the two-stage training approach for MSFFT.

3.2.3. Two-stage training approach

The MSFFT vehicle re-identification framework is developed to re-identify vehicles across a network of CCTVs and UAV. The network is exposed to learn the transformations of a vehicle observed from contrasting views. The bottleneck fused feature representation of vehicles generated by MSFFT is used further to train the network with self-supervised approaches namely SimSiam and NNCLR. Later the knowledge gained by the network after being trained using these approaches, is further utilized to perform the re-identification task in a supervised approach.

Training MSFFT Self-supervised SimSiam:

SimSiam [67] is a self-supervised method that aims to maximize the similarities between two augmented versions of an image. Instead of augmenting the images of vehicles observed by CCTV or UAV, identical vehicles observable by either surveillance platform (CCTV/UAV) are chosen as augmented versions of each other. This training method enables SimSiam to become less contrastive when re-identifying similar vehicles observed by CCTVs and UAVs. Like SimSiam, the encoder network $f(\theta)$ employs MSFFT as the backbone network with the projection MLP head. Given the MSFFT (Fig. 4), a fully connected layer with Batch Normalization (BN) is added after the final layer of MSFFT. For the two identical vehicles observed by cross platform, a prediction MLP head denoted as $p(\theta)$ aims to match the output of one view to the other. The prediction MLP shares the same architecture as SimSiam's self-supervised approach. The weights are shared between two encoders that process batches of identical vehicles of contrasting appearances. If x_{CCTV} and x_{UAV} are the two images of the same vehicle, the similarity between two contrasting views is maximized by minimizing the negative cosine similarity:

$$S(p_{CCTV}, z_{UAV}) = -\frac{p_{CCTV}}{\|p_{CCTV}\|_2} \cdot \frac{z_{UAV}}{\|z_{UAV}\|_2} \qquad (2)$$

where $p_{CCTV} = p(f(x_{CCTV}))$ and $z_{UAV} = f(x_{UAV})$ are the two output vectors of two different views (UAV & CCTV). Here $p(f(x_{CCTV}))$ is the vector obtained from the prediction MLP of CCTV vehicle image of MSFFT and $f(x_{UAV})$ is the vector generated by the projection MLP of UAV vehicle image of MSFFT.

SimSiam uses a symmetric loss using a stop-gradient operation where the gradient is propagated to an encoder from another encoder's prediction MLP head.

Training MSFFT using Self-supervised NNCLR:

NNCLR [68] self-supervised approach maintains a support set that includes embedded images from a subset of images in the dataset to facilitate the generation of new data points. NNCLR uses contrastive loss (in this case, InfoNCE loss) for discriminating between instances. For any two sample z_i and z_i^+ where z_i^+ is the augmented version of z_i, the InfoNCE loss is defined as:

$$L_{InfoNCE} = -log \frac{exp(z_i . z_i^+ / \tau)}{exp(z_i . z_i^+ / \tau) + \sum_{z^- \in \mathcal{N}} exp(z_i . z_i^- / \tau)} \qquad (3)$$

where τ is the softmax temperature and (z_i, z_i^+) (z_i, z_i^-) are the positive and negative pair. MSFFT is the backbone architecture for the encoder. The projection MLP contains three fully connected layers of sizes $[2048, 2048, d]$. Here d represents the total identical vehicles observed during training. The prediction MLP consists of two fully connected layers of size $[2048, d]$. Given a minibatch of identical vehicles $\{x_{ic}\}$, $i \in V_{id} = \{1, 2, 3, 4 \ldots .M\}$ and $c \in C_{id} = \{1, 2, 3, 4, \ldots, N\}$, the identical vehicles observed at different cameras (CCTV's) are sampled and fed into the encoder network $f(\theta)$. Here V_{id} is the identical vehicle observed by a network of cameras and a UAV denoted by the set C_{id}. The positive pairs (z_i, z_i^+) are obtained by identifying the nearest neighbor of $z_i's$ in the support set Q_{UAV}. The support set contains samples of identical vehicles observed by another platform (UAV). The NNCLR loss is defined as:

$$L_{NNCLR} = -log \frac{exp(NN(Z_i, Q_{UAV}) . Z_i^+ / \tau)}{\sum_{t=1}^{n} exp(NN(Z_i, Q_{UAV}) . Z_t^+ / \tau)} \qquad (4)$$

Here $NN(z, Q_{UAV})$ is the nearest neighbor operation for a CCTV view of the vehicle present in the support set. Note that each self-supervised approach is trained independently and further used for re-identification.

Training MSFFT using supervised approach:

Using the self-supervised approach the features learned by the backbone encoder network MSFFT are further utilized to conduct re-identification by training the final layers of MSFFT using self-supervised learning. Specifically, the weights of the encoder network (MSFFT) are frozen until the second level of the multi-scale feature extractor module of MSFFT and the last level of the multi-scale feature extractor module together with the fully connected layers are trained in a supervised approach with triplet loss.

4. Experiments

In the present study, a novel re-identification framework is proposed to perform cross platform vehicle re-identification. The re-identification network is evaluated on the developed dataset MCU-VReID. For the 51 annotated identical vehicles observed across 42 CCTVs and a UAV, for training the re-identification network, a total of 29 vehicle identities were obtained. The training set comprises 2366 vehicle images of 29 different vehicle identities. In the inference phase, the gallery/test consists of 3264 vehicle images of the remaining 22 vehicle identities. For each identical vehicle considered in inference, the vehicle images captured by both UAV and random surveillance cameras are considered as query vehicle images. This yields a total of 44 query vehicles. All the vehicle images used for experiments are resized to a uniform dimension of 224 × 224. Providing manually annotated training datasets of sufficient size is a challenge in supervised deep learning methods [70]. To tackle this issue, the underlying CNN layers of MSFFT are trained on a large dataset, and then the learned weights are used to perform the required task at hand. Before modeling the MSFFT, the convolutional layers (the initial layers, inception block) are trained without the transformer network with the standard vehicle re-identification dataset VeRI-776. The final concatenated convolutional features are flattened and are passed to the two fully connected layers with a Dropout of 0.5 between each layer. The final fully connected layer is used to predict the class labels. The network is trained using sparse categorical cross-entropy loss with an Adam optimizer. To perform transfer learning, the VeRI-776 training set that comprises 37,746 images of 769 vehicle identities is utilized. Further, the network is built upon by freezing the learned weights and adding the transformer network at each level of MSFFT to capture features at different scales.

The Section 4.1 outlines the criteria for choosing the T_{shot} value, which is subsequently used for generating keyframes for CCTV and UAV videos. Further Section 4.2 provides the implementation details of MSFFT used to carry out cross platform vehicle re-identification. To evaluate the performance of the network, available re-identification metrics such as mAP and rank-k accuracy are considered. The Section 4.4 presents the evaluation of the developed dataset with state-of-the-art re-identification methods. Table 3 summarizes the overall mAP and rank-k accuracy for re-identification methods evaluated on MCU-VReID. In order to train the re-identification models, Intel's Intel Xeon W-215 CPU @3.70 GHz was used along with 32 GB of RAM and an Nvidia Quadro RTX 4000 GPU has been used.

4.1. Estimation of T_{shot} using shot boundary

The shot boundary is detected by comparing the histogram variation between two successive frames using a threshold value, T_{shot} 3.1.1. The present study utilizes this technique for two separate surveillance platform systems that vary in the method of video acquisition. The CCTV cameras are static and have a fixed field of view. During surveillance if there are no major traffic movements, the background scene does not change very significantly from frame to frame. However, there is a notable change in the scene when there is an appearance or movement of an object. The data collected by CCTVs contain non-uniform vehicle movements where some cameras monitor significant vehicle traffic (Cameras located inside the campus) while the rest contain minimal traffic movements (Cameras located at hostel premises). In UAV videos there is a constant change in the scene due to the flight movement. Hence even if there is no sudden appearance or movements of entities in the scene, there is a drastic change in scene information due to the UAVs motion. The study aims to reduce the presence of static frames within a sequence of frames, where there is minimal change in the scene. Fig. 7 illustrates the histogram difference calculated for three distinct CCTV videos and a UAV video captured at various locations. The occurrence of a peak at regular intervals indicates the existence of a substantial change between two successive frames. As shown in Fig. 7, the value for T_{shot} is experimentally determined to be 0.2 for CCTV and 0.1 for UAV videos.

4.2. Vehicle re-identification using MSFFT

MSFFT employs a two-stage training strategy to re-identify vehicles across cross platform surveillance systems. At the first stage a self-supervised training is performed to learn various transformation views of vehicle images, as observed by both CCTVs and UAVs, in the cross platform vehicle re-identification process. Further at second stage of training, the knowledge about the vehicle transformation learned at first stage is utilized to train the MSFFT in a supervised manner for vehicle re-identification. As a baseline approach, single-stage training using supervised learning with triplet loss is used to perform re-identification with MSFFT. The hyperparameters that are used to train MSFFT in a supervised manner for both approaches (baseline and two-stage training) are kept in common.

Baseline implementation of MSFFT with supervised training:

The re-identification network inputs batches of identical vehicles observed by both CCTVs and UAV. The network mines the $P * K$ batches of images using the Batch Hard Triplet Loss variant with a soft margin [71]. Here K instances of P vehicle identities are sampled at each iteration in the course of training the re-identification network. P and K are experimentally set to 8 and 4, respectively.

The batch of $P * K$ images of dimension 224 × 224 are fed into the initial layers of MSFFT, which learns the initial contextual information of the vehicles. The first concatenated feature map is passed to the transformer blocks with a patch size of 16 × 16. Similarly, the features generated at successive inception layers are propagated to its transformer block to calculate different levels of attention scores. One key finding is that the number of patches obtained at each level of transformer blocks is 49 with dimensions of 7×7. However, each of the transformer blocks encodes the attention of the convolution feature maps at different resolutions. A trainable linear projection translates the flattened 2D patches to a constant latent space dimension D. The parameter D is experimentally chosen to be 64. The MSFFT with selected parameter D has fewer trainable parameters, making it lighter and shallower than ViT. The attention scores of each flattened patch at different transformer blocks are added to obtain the scores at different scales.

The re-identification network is trained for 500 epochs using Adam optimizer with a learning rate of $1e{-}4$. During inference, the final dense layer of the MSFFT is used to extract the feature for computing mAP and rank-k accuracy. A k-reciprocal re-ranking [72] mechanism is adopted to refine the initial ranking results of the re-identification network for a given query in order to optimize its performance. From Table 3 it can be seen that MSFFT trained using a supervised approach yields a re-identification score of 23.09% which is significantly less as compared with MSFFT with two-stage training approaches.

In self-supervised approaches, the image is enhanced to make the model invariant to semantic transformation. Self-supervised models are less invariant for handling inputs that have undergone heavy transformations for downstream tasks. Commonly used data augmentation techniques are flipping images, rotation of images at different angles, random resize crop, image blurring, etc. However, certain augmentation techniques, such as flipping and rotation, may change the semantic information. Rather than flipping and rotating the vehicle images observed through various cameras and UAV, the batches of input vehicle images to the network are considered to be transformed versions of one another. Along with this, other augmentations such as brightness, contrast, saturation, and hue are applied with the strength of [0.4,0.4,0.4,0.1] along with random resized cropping of scale 0.2 and 0.1 respectively. These set of data augmentation parameters are kept common to train MSFFT in a self-supervised manner as a part of a two-stage approach for vehicle re-identification.

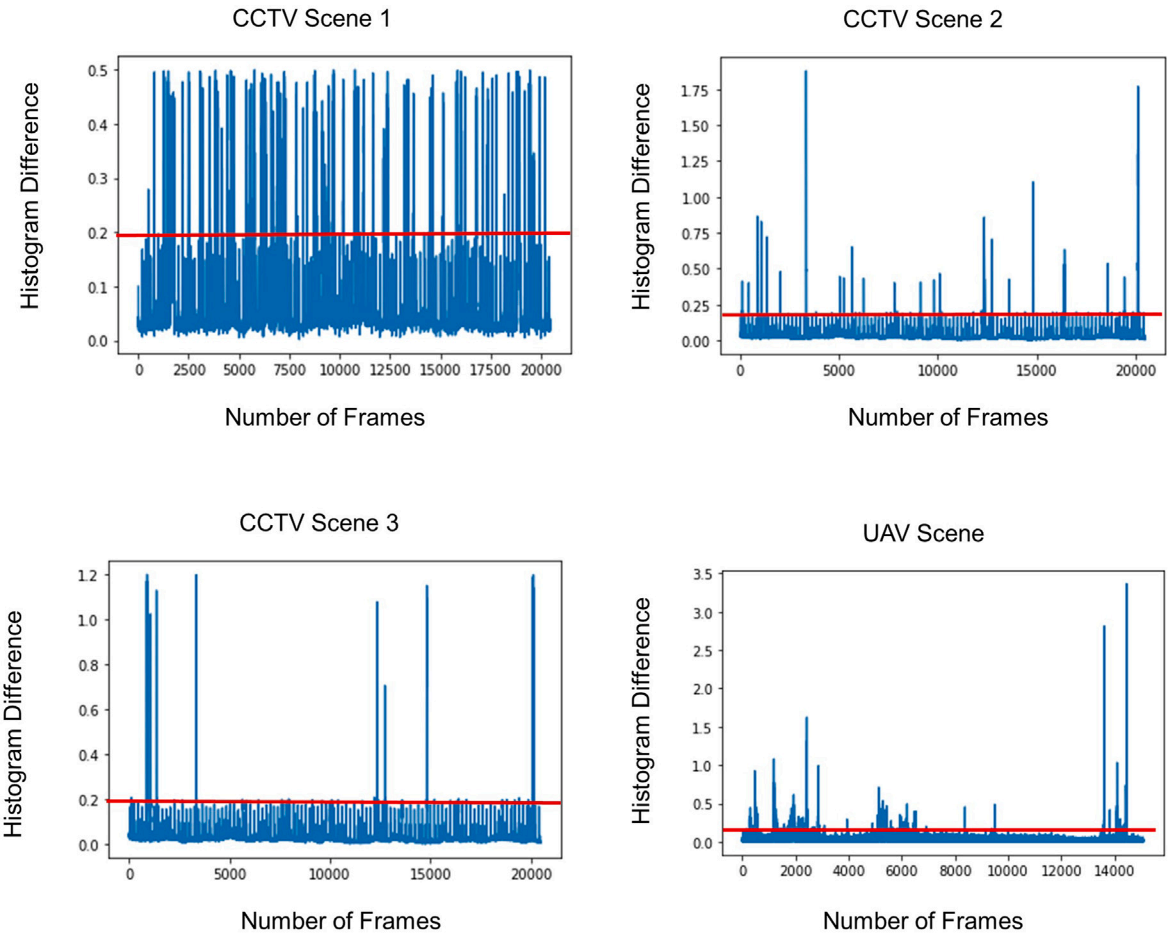


Fig. 7. Variation in Histogram difference for videos of CCTVs and UAV.

Table 3
Comparison with state-of-the art re-identification frameworks on MCU-VReID.

Method	Training strategy	mAP	rank-1	rank-5	rank-10	rank-20
Single-stage Training: Supervised Learning (SL), Unsupervised Learning (USL) and Self-Supervised Learning (SSL)						
PCB-4 [23]	SL	16.3	36.4	38.62	38.62	43.2
PCB-6 [23]	SL	17.5	29.5	36.4	43.2	47.7
OSNet [24]	SL	21.4	36.4	45	47.7	52.3
ICE [25]	USL	18.71	38.6	43.18	50	50
PPLR [26]	USL	18.8	34.1	43.2	47.7	50
Nformer [27]	SL	18.2	31.8	40.9	52.3	56.8
PAT [28]	SL	20.74	31.8	40.9	47.7	52.3
AdaSP [35]	SL	21.1	40.91	45.45	50	52.27
ResNet-50 [31]	SL	18.4	27.3	40.9	44.2	50
ResNet-50 [31]	SSL (SimSiam [67])	20.8	28.8	43.25	47.7	50
ResNet-50 [31]	SSL (NNCLR [68])	21.9	36.36	45.45	48.6	52.2
TransReID [32]	SL	26.86	38.36	40.5	55.6	60.34
TransReID [32]	SSL (SimSiam [67])	24.26	28.8	42.4	47.7	56.26
TransReID [32]	SSL (NNCLR [68])	25.4	32.4	47.2	52.27	58.4
MSFFT (Proposed)	SL	23.09	36.36	45.45	50.09	63.36
MSFFT (Proposed)	SSL (SimSiam [67])	25.42	30.2	47.7	52.3	66.4
MSFFT (Proposed)	SSL (NNCLR [68])	26.26	38.36	45.45	50	62.4
Two-stage Training: Self-Supervised Learning (SSL) + Supervised Learning (SL)						
ResNet-50 [31]	SSL (SimSiam [67]) + SL	22	31.81	45.45	47.72	54.54
ResNet-50 [31]	SSL (NNCLR [68]) + SL	24.76	38.36	50	52.27	54.54
TransReID [32]	SSL (SimSiam [67]) + SL	27.1	34.4	45.45	52.2	60.4
TransReID [32]	SSL (NNCLR [68]) + SL	28.4	36.36	47.42	50	60.4
MSFFT (Proposed)	SSL (SimSiam [67]) + SL	25.49	31.81	50	54.54	68.18
MSFFT (Proposed)	SSL (NNCLR [68]) + SL	**29.76**	40.9	45.45	50	63.63

4.3. Implementation of MSFFT with two-stage training

To re-identify vehicles observed across cross platform surveillance systems, the underlying re-identification network must be robust enough to re-identify vehicles captured by two contrasting surveillance systems. MSFFT is trained with a two-stage training approach, where the network is made to learn different transformations using self-supervised approaches with SimSiam and NNCLR. Further, the re-identification is performed in a supervised approach for vehicle re-identification. The implementation details of both approaches for two-stage training are given below.

Two-stage training: SimSiam+Triplet Loss:

The MSFFT network identifies the semantic aspects of vehicles from different perspectives with a SimSiam approach. The network utilizes the training images without any labels from MCU-VReID, where the batches of identical vehicles from two different surveillance systems are fed to the encoder network separately. The network is trained for 200 epochs using SGD optimizer with a learning rate of $1e{-}4$ and momentum of 0.6. Negative cosine similarity is used as a loss function, which aims to maximize the similarity of identical vehicles observed across cross platform surveillance systems. The learned weights are further applied to perform vehicle re-identification following the same principle that is used to train baseline MSFFT. For supervised training, weights of the final level of the multi-scale feature module of MSFFT and the deeper dense layers are finetuned. The same training images are used with labels for training the network with triplet loss. From Table 3 it can be seen that MSFFT re-identification with SimSiam self-supervised approach achieves an mAP of 25.49% that is significantly higher than the supervised approach of re-identification with MSFFT.

The base implementation of SimSiam with ResNet-50 backbone is also modeled to perform re-identification. A pre-trained ResNet-50 with ImageNet weights is used as a backbone encoder network. The ResNet-50 comprises five residual layers with skip connections. The feature map of the last residual block is pooled and fed to the final fully connected layers for performing any vision task. For self-supervised learning, the last residual block is fine-tuned, and a similar approach to that of MSFFT is adopted to train the network. The self-supervised approach is then applied to train the re-identification network. During inference, an mAP of 22.0% is achieved.

Two-stage training: NNCLR+Triplet Loss:

In a similar manner to SimSiam self-supervised MSFFT re-identification, the training images of MCU-VReID are selected to train MSFFT with NNCLR. Here the labeled identical vehicles observed across different CCTV cameras are fed into the encoder (MSFFT). The support set contains batches of unlabeled identical vehicles observed by UAV. The support set is a queue updated at every iteration of the training step. The support set size is experimentally set to 50. For a given embedding of an encoder of a particular CCTV view, its nearest neighbor from the support set is retrieved, and InfoNCE loss is calculated. The network is trained for 200 epochs with Adam optimizer at a learning rate of $1e{-}4$. The learned encoder network is used to train the re-identification network by finetuning the final layers of MSFFT. From Table 3 the mAP for MSFFT with NNCLR approach yields an mAP score of 29.76%, which is significantly higher than other re-identification models. The NNCLR approach is also explored by using ResNet-50 as the backbone encoder network, similar to the SimSiam experiment of performing re-identification. The inference of re-identification yields an overall mAP of 24.76%. The increase in mAP highlights the robustness of MSFFT for re-identifying vehicles across multiple scales over ResNet-50 for cross platform vehicle re-identification.

4.4. Experiments with state-of-the-art re-identification networks

The performance of the proposed vehicle re-identification framework is validated against the state-of-the-art re-identification methods. As discussed in Section 3.1, the existing vehicle re-identification methods utilize extra supplement information of vehicles to carry out re-identification. The developed MCU-VReID dataset does not provide this information apart from the vehicle and camera identity. In contrast, the person re-identification methods utilize very little supplementary information to perform re-identification. Different person re-identification also shares the same principle to perform re-identification. Hence in this work, the proposed vehicle re-identification framework is compared with the state-of-the-art person re-identification (Section 2.2.1) and a transformer-based object re-identification method (Section 2.2.2) with little supplementary information of vehicles. Note that to the best of our knowledge, there is no existing work on vehicle re-identification using two different surveillance platforms (CCTV and UAV).

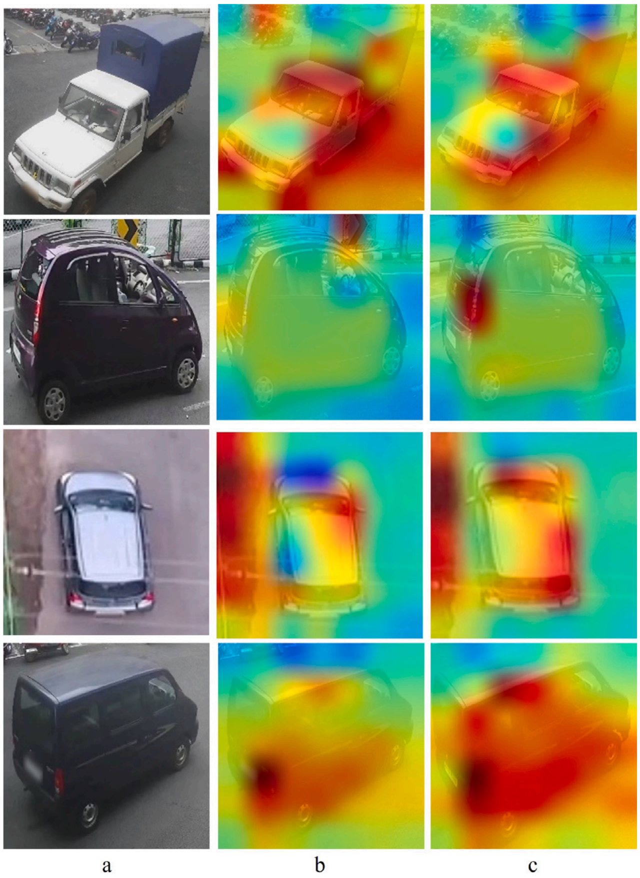


Fig. 8. Comparison between the generated heatmap for MSFFT using SimSiam and NNCLR approaches. For each example (a), the Heatmap obtained using SimSiam (b) and NNCLR (c).

The developed dataset experiments with PCB [23], which produces the convolutional feature descriptors that represent part-level features. The PCB network is trained for two different values of ‘P’, i.e., P = 4 and P = 6, that define the number of uniform partitions to be performed on the convolutional feature descriptor. As a backbone network, PCB makes use of the standard ResNet-50, which is initialized with ImageNet weights. The network is trained using triplet loss by concatenating the feature descriptor of each part across the total number of convolutional channels. The network is trained with an SGD optimizer for a learning rate of 0.001 and weight decay of 5e−4 for 5K iterations. For each of the experiments, i.e., PCB-4 and PCB-6, the re-identification scores are computed with an mAP of 16.3% and 17.5%, respectively. It can be seen from Table 3 that the performance of PCB re-identification is significantly lower than the proposed framework for vehicle re-identification.

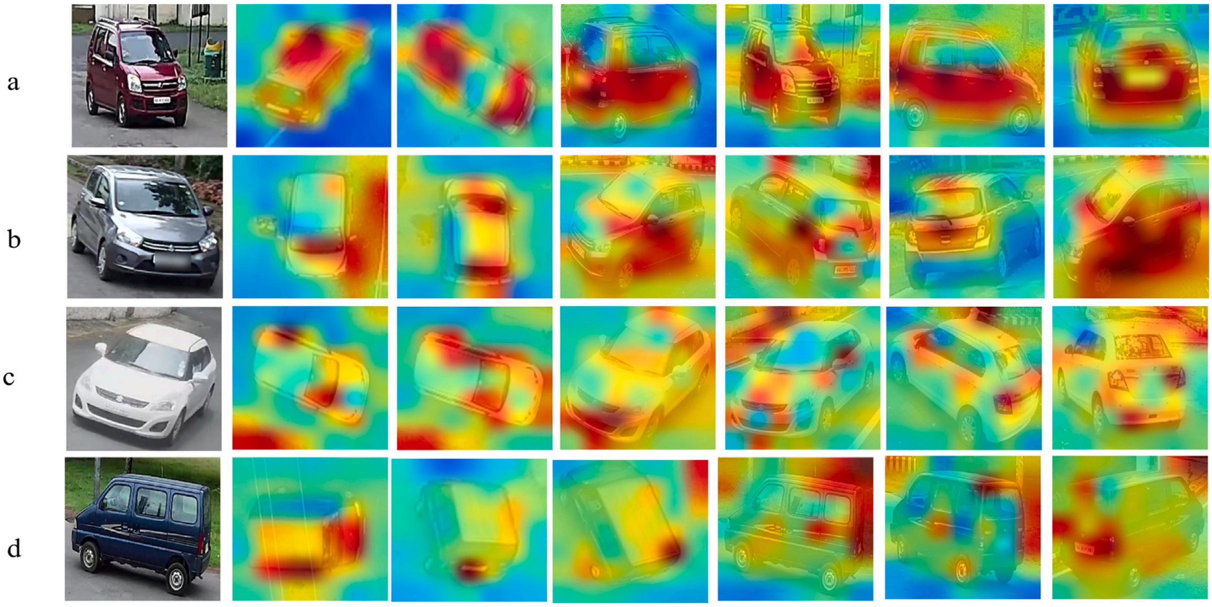


Fig. 9. Illustration of heatmaps generated by MSFFT NNCLR approach for vehicles observed by cross platform surveillance systems.

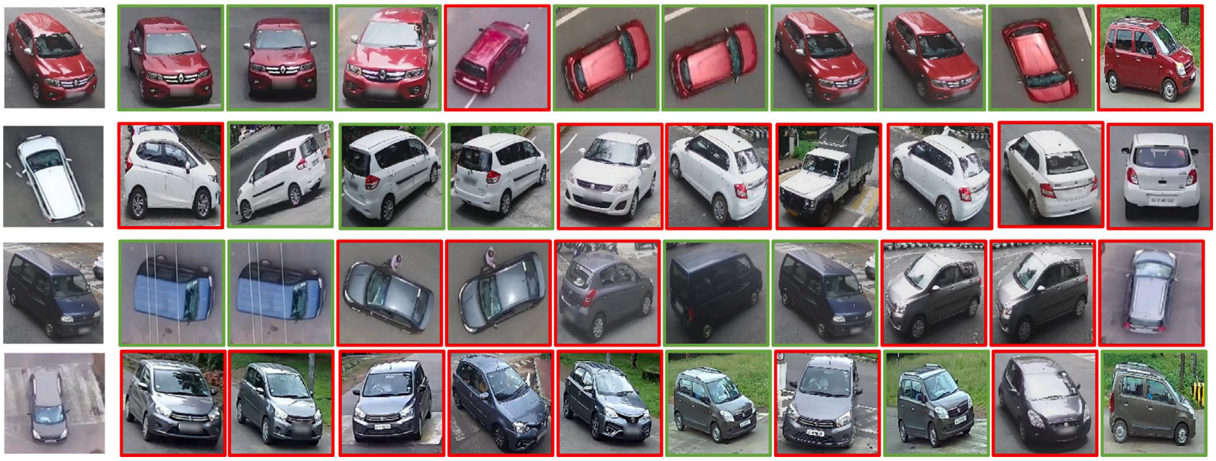


Fig. 10. The retrieval of the top 10 identical vehicles for a given query. At each row, the first image is the query for which the top-10 identical vehicles observed across CCTVs and UAV are retrieved. A retrieved identical vehicle with a green border is a positive match, and a vehicle with a red border indicates a vehicle that has been incorrectly re-identified.

Table 4
mAP and rank-k accuracy for individual surveillance platform vehicle images of MCU-VReID.

Method	Surveillance platform	Approach	mAP	rank-1	rank-5	rank-10	rank-20
MSFFT (Proposed)	CCTV	SL	26.03	38.8	50	55.5	77.7
MSFFT (Proposed)	UAV	SL	23.1	34.09	45.45	50	56.81
MSFFT (Proposed)	CCTV	SSL (NNCLR [68]) + SL	26.46	44.4	50	55.5	61.1
MSFFT (Proposed)	UAV	SSL (NNCLR [68]) + SL	23.22	45.45	50	63.63	63.63

For TransReID [32] vehicle re-identification on the MCU-VReID dataset, a pre-trained ViT-B/16 with a step size of 16 is chosen as a backbone feature extractor network. The pretrained TransReID uses camera and viewpoint information from VeRI-776 for re-identification. To evaluate the performance of MCU-VReID the network is finetuned for the final layers of ViT-B. The network is trained using an SGD optimizer with a weight decay of 1e−4. The base learning rate of 0.008 is chosen to train the network for 500 epochs. From Table 3 it can be observed that an mAP of 26.86% is achieved for the MCU-VReID dataset.

The developed MCU-VReID dataset is also evaluated with recent frameworks for re-identification namely ICE [13], NFormer [15], PPLR [14] PAT [16] and AdaSP [17]. From the Table 3 it can be seen that these frameworks for vehicle re-identification yields a low mAP when compared to the proposed approach. Section 5 provides an explanation on the performance of each experiments presented in Table 3.

Fig. 8 shows the generated heatmap for using the re-identification net with SimSiam and NNCLR approach. With both techniques, the network is able to locate key points on the vehicle. Compared to SimSiam, MSFFT trained with NNCLR is more likely to localize more keypoints. Fig. 9 illustrates the heatmap obtained for identical vehicles using MSFFT re-identification net with NNCLR approach. With the vehicles observed at different cameras, the model is able to locate keypoints of the vehicle irrespective of dynamic perspectives and scale. CCTV and UAV views should share overlapping information on vehicle keypoints, so that similarity can be estimated between the two views.

Fig. 10 shows the top-10 identical vehicles retrieved from the gallery for each query example. Even with severe appearance changes of vehicles, MSFFT trained with NNCLR can find the most similar vehicle similar to the given query in the top-10 list. Examples (b) and (c) in Fig. 10 contain a query vehicle identity observed by UAV, and the top-10 retrieved vehicle images from the gallery correspond to vehicle identities observed by CCTV cameras. This is due to the objective of conducting inter-camera re-identification rather than intra-camera re-identification.

The Table 4 presents the re-identification scores for MCU-VReID when considering the vehicle images observed from standalone CCTVs and UAV surveillance cameras. Two experiments are conducted with vehicle images belonging to individual surveillance systems. An mAP of 26.03% and 23.1% is achieved for CCTV and UAV vehicle re-identification. Two-stage training is also adopted for platform specific vehicle images where an mAP of 26.46% and 23.22% is achieved for re-identifying vehicles from CCTVs and UAV platforms respectively.

5. Discussion

This section presents a discussion of MCU-VReID's performance on the existing and the developed re-identification model.

MCU-VReID when evaluated on the existing re-identification models, the performance was found to be significantly poor. PCBNet, designed for person re-identification, focuses on prominent parts of the person's body by stripping the convolutional features to 'p' horizontal strips. The location of parts of a person, when observed from different perspectives, ensures that positional information remains intact. For vehicle re-identification, the semantic information of vehicles differs significantly. Additionally, cross platform vehicle re-identification contains vehicle data from aerial and side perspectives, which can mislead in justifying exact parts information while stripping convolutional features. OSNet addresses the scale invariance issues associated with objects observed on arbitrarily large scales. However, MCU-VReID contains the vehicle observed from two contrasting views taken by UAV and CCTV.

Vehicle re-identification using ICE [25] relies on an unsupervised training strategy for generating pseudo-labels for vehicle instances. Due to the significant change in the distribution of vehicle instances observed by two contrasting surveillance systems, ICE attains a lower re-identification score of 18.71%.

Vehicle re-identification using the PPLR [26] approach yielded a mAP of 18.8%. PPLR attempts to learn semantic information by understanding the complementary relationships between global and local part features. Since PPLR assumes that the data distribution of entity instances belongs to the same distribution, modeling the cross-view-based vehicle re-identification resulted in generating indistinct pseudo labels.

Re-identification of vehicles using NFormer [27] yielded a mAP of 18.2% which is significantly lower than the proposed approach. Re-identification using the NFormer approach relies on learning a robust global representation of vehicles by modeling the interaction of vehicle instances using a relation map. NFormer approach is expensive in computation as it models these interactions with every vehicle instance appearing in a mini-batch. Further due to drastic changes in vehicle dimension, appearance, and viewpoint the NFormer based approach for re-identifying vehicles is highly prone to isolate certain instances as outliers.

Vehicle re-identification using PAT [28] attained poor re-identification yielding a mAP of 20.74%. In PAT the backbone feature extractor follows a similar approach to PCBNet for learning semantic local features by using three tokens for understanding the structure of person geometry. However, for vehicle re-identification where there is a significant change in viewpoints and appearance, the PAT using these patch tokens fails to understand the semantic structure of vehicles resulting in poor re-identification.

Using Adaptive Sparse Pairwise Loss (AdaSP) [35] for vehicle re-identification achieved a mAP of 21.1%. Since AdaSP online mines the top-k similar positive vehicle instance pairs, it fails to sample adequate positive pairs of vehicle instances that have significant viewpoint and appearance changes.

Transformer-based re-identification using the TransReID approach produced comparable results to PCBNet and OSNet. Through the use of ViT-B/16 as a feature extractor, TransReID can learn more local features of vehicles without losing any information than with CNN-based approaches. However, it applies a fixed patch size across blocks of multi-head attention to computing attention scores. TransReID also uses learnable parameters to learn the embeddings for viewpoints and camera information which requires an indirect way of aiding re-identification by providing additional annotations. TransReID yields better re-identification accuracy of 26.86% as compared to the baseline MSFFT with a re-identification accuracy of 23.09%. To evaluate TransReID on the developed dataset, a pre-trained model of ViT-B/16 [34] is chosen. The model was initially trained on VeRi-776 [19] that utilized additional viewpoint information to compute Side Information Embeddings (SIE). The VeRi-776 [19] provides additional supplemental information regarding the viewpoint labels for enhancing the training of re-identification. However, in the present study, vehicle re-identification is performed across a network of CCTVs and UAV with no additional supplementary information. In comparison to VeRi-776 dataset, the present study contains vehicles observed by both CCTVs and UAV that exhibit significant viewpoint changes and scale. Providing/Annotating these vehicles with viewpoint information is challenging as the semantic information poses confounding situations which may lead to inaccurate labeling. To validate whether the inclusion of the extra supplemental information have aided the performance of TransReID over MSFFT in supervised training, the TransReID is modeled to train with self-supervised learning and two-stage training approaches. From Table 3 it can be seen that TransReID trained with two-stage training approach with SimSiam and NNCLR results in a re-identification score of 27.1% and 28.4% respectively which lower than the proposed framework for re-identification. For the experiment the TransReID with feature extractor ViT-B/16 is trained with MCU-VReID without any labels at first stage to learn the semantic perturbations of vehicles (SSL Training) and further fine-tuned for re-identification in a supervised approach using vehicle labels. It can be inferred from Table 3 that for single-stage training the ResNet-50 and the proposed method using self-supervised learning performs better than supervised learning. On contrary TransReID with supervised learning yielded a higher mAP than self-supervised learning. By training the MSFFT which includes the multi-scale feature extractor modules using a self-supervised approach, it is able to learn the various semantic information of vehicles required for cross platform re-identification. Hence, the proposed feature extractor framework designed to learn the transformation using two-stage training performs better than TransReID for cross platform vehicle re-identification. This validates the observation in considering the additional learnable parameters by TransReID resulted in obtaining a better mAP than MSFFT in supervised training. In Section 6, an analysis of the effectiveness of MSFFT over TransReID is presented while re-identifying the vehicles across challenging scenarios.

The developed vehicle re-identification framework is also evaluated for surveillance platform-specific vehicle images. From Table 4, two re-identification experiments are performed using MSFFT in a supervised approach. During each of these experiments, the MSFFT is trained and evaluated on vehicle images from MCU-VReID corresponding to each surveillance platform. It can be observed that the performance of vehicle re-identification in both standalone surveillance modes (26.03% mAP for CCTV and 23.1% for UAV) is lower than vehicle re-identification experiment (mAP of 29.76%) that is trained and tested with vehicle images acquired by both surveillance platform.

For each of the individual vehicle re-identification experiments presented in Table 4, the re-identification framework is trained using a subset of vehicle images from the MCU-VReID dataset that is specific to either CCTVs or UAV. This leads to fewer vehicle instances of an identical vehicle that is needed when training the network with triplet loss. To train the re-identification network, triplet loss requires more 'K' vehicle instances for each 'P' vehicle identity. In training the re-identification framework, especially for CCTV surveillance platform, the network will learn fewer features about a vehicle's identity if fewer images of the same vehicle are observed across other cameras. To evaluate the performance of CCTV re-identification, for every query image observed at a particular camera, the priority is to find/rank similar vehicles observed across the rest of the cameras. This is to encourage cross-camera re-identification. During UAV re-identification experiments (Table 4), vehicle images related to UAV are used to train the re-identification framework. However, compared to vehicle re-identification with CCTV cameras, the re-identification framework is evaluated for batches of vehicle images acquired by UAV. Thus, when considering a framework for a query vehicle image captured by UAV, the objective is to identify similar vehicle instances captured by the UAV when traversed at different locations.

It may be noted that from Table 3 that the performance of cross platform vehicle re-identification with MSFFT using a supervised approach with an mAP of 23.09% is significantly less than the experiments that are trained and evaluated on platform-specific vehicle images from MCU-VReID (Table 4). cross platform vehicle re-identification using MSFFT is trained on batches of images that comprise vehicle images seen in two contrasting views (CCTVs & UAV). These images, when trained with a supervised approach, fail to generalize the representation of vehicles observed across two different views. In contrast, the network that is trained with platform-specific vehicle images for experiments conducted in Table 4 can generalize the representation of vehicles as they are trained and evaluated with platform-specific vehicle images of MCU-VReID. This study assumes that the vehicles observed by two contrasting surveillance systems have two distinct feature spaces for vehicle images seen by CCTVs and UAV. Thus the two-stage training is also used to train the network with platform specific vehicle images. As the training stage inputs batches of platform specific vehicle images, additional data augmentation in the form of horizontal and vertical flipping is applied. Data augmentation is applied to self-supervised learning to examine whether the performance of re-identification increases. It can be noted from Table 4 the re-identification accuracy increases slightly by 26.46% and 23.33% for CCTV and UAV vehicle re-identification.

Therefore, for cross platform vehicle re-identification, a self-supervised approach is necessary to discover the similarities between the two contrasting feature spaces before performing re-identification in a supervised manner. This can be seen from Table 3 that training the network initially in a self-supervised way and further utilizing it to perform re-identification in a supervised manner yields a better re-identification score of mAP of 29.76% that is higher than the re-identification scores for the experiments tabulated in Table 4.

All the above approaches are trained using supervised approaches with triplet loss. In existing vehicle re-identification methods, vehicle images are acquired by standalone surveillance systems (CCTV/UAV). Hence, with a supervised approach to re-identification, this study assumes that the vehicle images observed by either CCTV or UAV have their own representative feature space. Training the network in a supervised approach for two contrasting views of a vehicle for re-identification may yield a poor score. This is evident for the baseline MSFFT where a re-identification score of 23.09% is obtained, significantly less than the method learned using a self-supervised approach with further supervised learning using triplet loss. Rather than training the model initially with a mixture of vehicle images from two contrasting surveillance systems, the re-identification framework is exposed to learn the features of vehicles appearing in two contrasting views. Further, using these learned representations, the network is fine-tuned slightly to perform re-identification in a supervised approach.

6. Ablation study

To analyze the effectiveness of the proposed method the study conducted extensive experiments for comparison with state-of-the art re-identification frameworks. In Section 6.1, the performance of these re-identification frameworks is validated on the developed dataset with two different training strategies namely single-stage and two-stage training. Further in Section 6.2 the study highlights the significance of multi-level feature concatenation of MSFFT for re-identifying the vehicles appearing in cross-platform surveillance systems. The study also presents an analysis of how the concatenated feature of MSFFT feature extractor is beneficial for cross platform vehicle re-identification. Finally, the study showcases how MSFFT performs better than TransReID in challenging conditions for querying vehicles in certain scenarios.

6.1. Effectiveness of two-stage training strategy over single-stage training on MCU-VReID

The study analyzes the necessity of adopting the two-stage training for re-identifying the vehicles as seen by cross platform surveillances systems. It can be observed from Table 3 that the performance of re-identification methods is categorized under two different training strategies. For single-stage training approaches, the re-identification learns the vehicle features using supervised learning, unsupervised learning and Self-supervised learning. Frameworks that are modeled to train either with supervised and unsupervised learning yields a lower mAP re-identification scores than Self-supervised learning. Supervised learning mines the batches of vehicle instances to train the model for distinguishing the positive and negative class of identical vehicles. However, these vehicles observed by two different surveillance systems, exhibit severe appearance changes making the training process less effective for learning semantic information of vehicles. The frameworks presented in Table 3 which are modeled to train using unsupervised approaches generates pseudo-labels or instance-contrastive loss for learning the semantic features of vehicles. It can be observed from Table 3 that the performance of the re-identification models using unsupervised approach is comparable with supervised learning. Re-identification models that are trained using self-supervised approaches namely SimSiam and NNCLR yields a better re-identification scores than supervised/unsupervised approaches. Models that are trained using these approaches learns semantic perturbations of vehicles by mining camera specific (UAV/CCTV) vehicle instances or determining the nearest-neighbors of each other for estimating the similarities. Thus models trained using these approaches are able to learn two semantically different vehicle distributions effectively.

Since single-stage training approach using self-supervised learning results in a better re-identification accuracy, this study adopts self-supervised strategy to learn the semantic transformations of vehicles seen observed across cross-platform surveillance systems. It can be seen that the re-identification frameworks when evaluated using two-stage training approaches yields a better re-identification scores over single-stage training. This validates the assumption made in the study that vehicles observed by the cross-platform surveillance cameras having significant viewpoint, appearance changes as they belong to different distribution of information. Thus for cross-platform based vehicle re-identification with minimal labeling and information about vehicle instances it is appropriate to consider two-stage training for re-identification of vehicles.

6.2. Effectiveness of MSFFT for cross-platform based vehicle re-identification

To evaluate the effectiveness of MSFFT for cross-platform based vehicle re-identification, two challenging scenarios of vehicle appearance is considered. Scenario 1 (Example (a) of Fig. 11) contains two identical vehicles with minimal overlapping parts/significant viewpoint changes

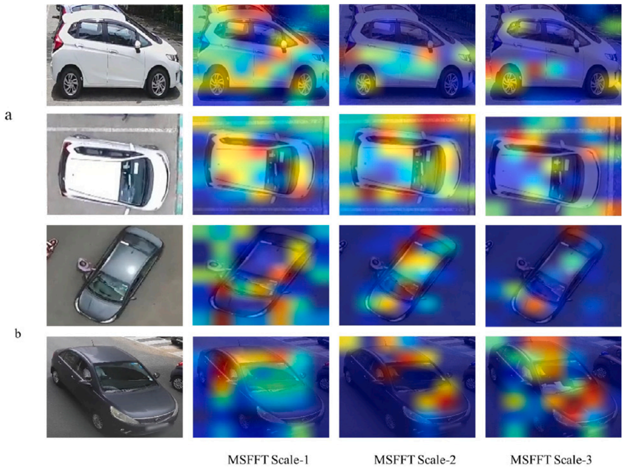


Fig. 11. Visualization of heatmaps generated by Grad-CAM [73] for two identical queries as seen by UAV and CCTV. The generated heatmaps are obtained from the final transformer layers at each level of MSFFT.

and Scenario 2 (Example (b) of Fig. 11) with subtle changes in the background and vehicle information.

The proposed feature extractor for cross-view-based vehicle re-identification learns vehicle features at three different levels. At each of these levels, the convolution operation that shares the design principles of inception layers aims to learn the semantic features of vehicles at different spatial locations with different kernel sizes. The learned representation or feature vector is further guided by transformer encoders to learn the semantic relationships of the vehicle parts by computing the attention score for a patch with the remaining patches of the feature map. As the convolution layers pay more importance to understanding the vehicle parts, the feature map generated by these layers only contains the learned representations of vehicle attributes. The transformer at each level encoder utilizes these feature maps where the patches containing vehicle parts or attributes have significant scores over others. At each level, the transformer encoder learns the semantic relationships of the vehicle features at different patch sizes owing to it learning more discriminative features of vehicles. Fig. 11 shows heatmaps generated by MSFFT at three different scales. In scenario 1 (Fig. 11), for the identical vehicles observed by both the platforms the feature extractor is able to learn the semantic information of vehicles that has a minimum overlap of vehicle parts. At the initial level, due to the larger patch size ($P = 16$ for Scale 1), the network aims to learn a more semantic representation of vehicles (vehicle roof, doors in Fig. 11). Further down the network with a decrease in patch sizes, the network aims to pay attention/give importance to local parts of vehicles (wheels of vehicles, side mirrors, headlights, etc. in Fig. 11). Utilizing the concatenated features learned from three different scales aids MSFFT with rich information of vehicles for better re-identification. For the identical vehicle pairs in scenario 2 (Fig. 11), at deeper level of MSFFT the network is able to pay less attention for background noise and focuses more on vehicle parts. The traditional vehicle re-identification approach would fail to recognize the vehicle due to the severe transformation between the two views. In contrast, the proposed framework MSFFT with two-stage training models the transformation between the two views and incorporates vehicle features at different spatial locations, and understands their semantic relationships through transformer encoder layers. Thus the features calculated at different levels effectively aid MSFFT for better cross platform vehicle re-identification.

6.3. Qualitative analysis of MSFFT and TransReID on MCU-VReID

A qualitative analysis is performed to evaluate the effectiveness of TransReID and MSFFT on MCU-VReID. For both the frameworks the two-stage training (SSl (NNCLR) + SL) approach resulted in a better re-identification than the rest. Hence the effectiveness of both TransReID and MSFFT is evaluated qualitatively using these experiments for re-identifying vehicles under complex situations. Fig. 12 presents two scenarios for re-identifying a query vehicles seen across cross platform surveillance systems. In the Fig. 12 for two query identities, using both TransReID and MSFFT a top-5 vehicles are retrieved from the gallery and their corresponding heatmaps using Grad-CAM [73] are generated. It can be observed that for the first query identical that shares a similar appearance with the background, both TransReID and MSFFT are able to get the candidate query vehicle instances in the top-5 retrieved information. However, for the top-5 retrieved information corresponding to TransReID, it can be observed that there is no vehicle instance of the query seen by CCTVs. It appears that TransReID prefers the top-5 retrieved image for a query from the same platform, as such a query image from CCTV will result in top-5 images of CCTV only.

For the generated heatmap using Grad-CAM [73] in Fig. 12, for a positive match of the query vehicle (Query 1 Fig. 12) in top-5, TransReID is not able to generalize the features of vehicles effectively from its background. On the contrary, the MSFFT can distinguish these subtle appearance changes and hence it is effective at paying importance to vehicle attributes. Addressing the second scenario where for a given query vehicle identity the TransReID framework for re-identification can retrieve a single surveillance camera specific (CCTV) vehicle instance from the gallery. Given a CCTV view of the query vehicle (Query 2 Fig. 12), TransReID is able to pay a marginal importance to a small overlapping portion of the positive match of the query. TransReID pays significant importance to the global appearance of vehicles with less attention to the local parts. This is due to the backbone feature extractor ViT-B/16 of TransReID that provides a feature representation of vehicles at a single scale. While MSFFT can retrieve those instances of query identity observed by both surveillance systems even with minimal overlap of vehicle parts. This presents an additional disadvantage for TransReID performance on MCU-VReID that when there is a lesser overlap of vehicle information due to significant viewpoint changes. On the other hand, MSFFT can pay importance to local contextual information about vehicles when there is a minimum overlap of vehicle information to estimate similarity. Based on the observations that were inferred from the experiments, the TransReID performs poorly in these scenarios (a): for a given query if it contains corresponding vehicle instances in the gallery with limited part overlap information. (b) the query and corresponding vehicle instances have similar appearance features with background.

Although the proposed re-identification framework outperforms the state-of-the-art re-identification methods, the re-identification is significantly less. This is due to the various challenges that are observed in the MCU-VReID dataset. Fig. 13 illustrates some of these scenarios. In Fig. 13, each row consists of vehicle identities observed across both UAV and CCTVs. For example (a) of Fig. 13, the vehicle has a dual-color, which makes re-identification challenging by both UAVs in aerial view and side views as seen from surveillance cameras. During data acquisition, there were scenarios in which the UAV was held still in the air that captured largely the aerial view of the car. This limits an inference of any additional parts information to perform re-identification with vehicles observed at different viewpoints by surveillance cameras. In the second example (b) from Fig. 13 i.e., pickup constitutes multiple colors and appearances, making re-identification difficult. In the last two examples (c) and (d) from Fig. 13, it can be observed that the vehicles that are acquired by UAV are of different scales that make the re-identification network just infer global features of the vehicle (color, edges). Owing to recording vehicles by UAV which should be seen in surveillance cameras, in specific scenarios, due to a non-uniform flight speed in some frames, the vehicle appeared to be blurry (First image in example (d)). Due to this, the re-identification becomes challenging and results in fewer scores.

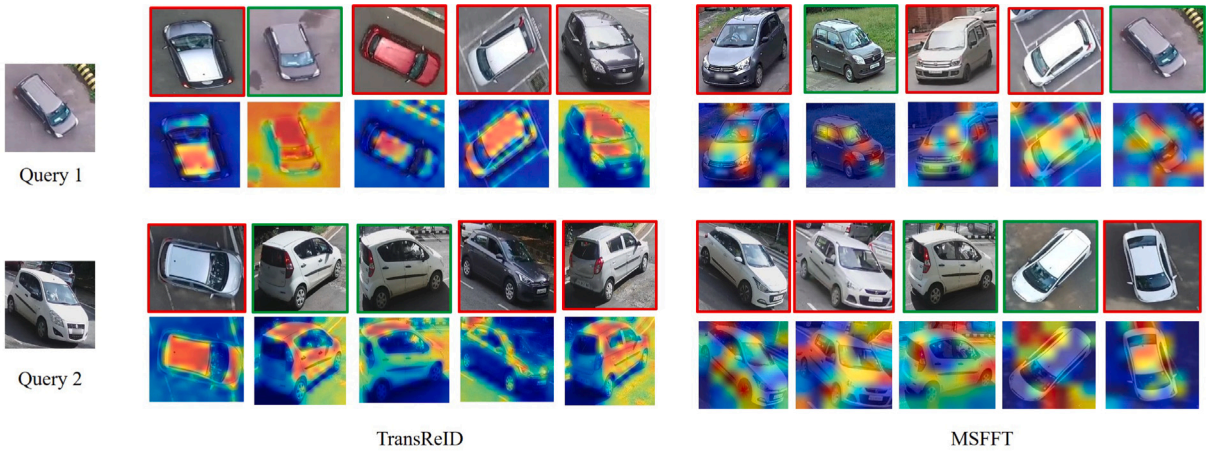


Fig. 12. Evaluating the effectiveness of TransReID with the proposed MSFFT framework for two scenarios of query vehicle images. For each of the two framework the top-5 vehicle instances and their corresponding heatmaps obtained from Grad-CAM [73] are illustrated.

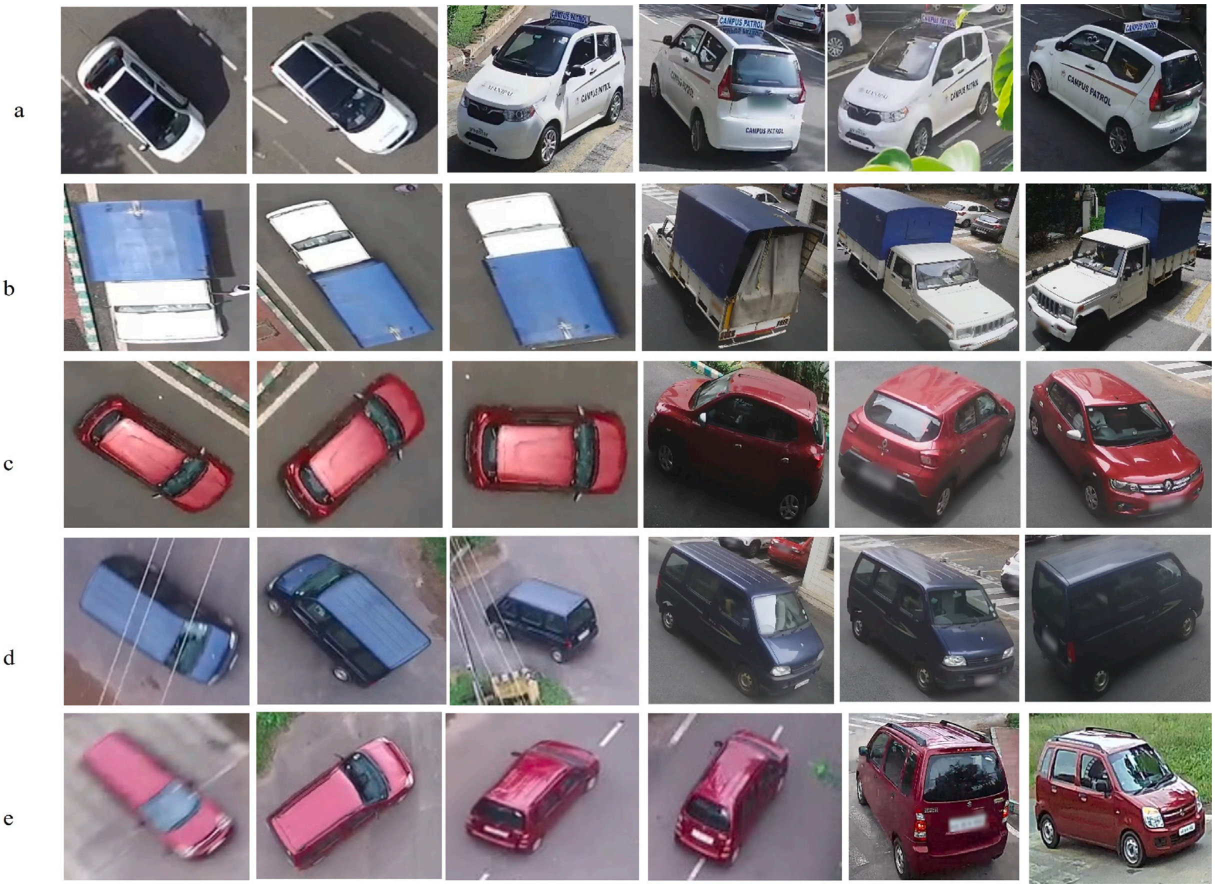


Fig. 13. Examples (a) and (b) illustrate the situation of a vehicle that has a subtle change in appearance and dimension changes. Example (c) shows a vehicle subjected to illumination changes along with the additional background. Examples (d) showcase vehicles taken at a dynamic scale by UAV and due to the fast movement of UAV, it leads to obtaining obscure vehicle images as shown in example (e).

7. Conclusion

Vehicle re-identification is an essential aspect of ITS in traffic monitoring. Earlier, re-identification is carried out using standalone surveillance systems such as CCTVs and UAVs. In this article, a novel vehicle re-identification dataset is developed to re-identify the vehicles observed across CCTVs and a UAV. The dataset comprises surveillance footage acquired from 42 surveillance cameras and a UAV at different locations on a gated university campus. For 51 identical vehicles observed across CCTVs and UAV a finer annotation is provided. A novel re-identification method is proposed to evaluate the developed dataset's performance. The proposed re-identification framework learns the semantic attributes and structures of vehicles at different spatial locations using inception layers and understands their semantic relationships by integrating the semantic feature maps with transformer layers

The proposed approach employs a two-stage training strategy in which the method is exposed to learn various semantic transformations through self-supervised learning methods. Using the knowledge gained by training in a self-supervised learning approach, the framework is finetuned to re-identify vehicles observed across cross platform surveillance cameras using supervised learning. The study validates the effectiveness of the proposed re-identification framework and the necessity of two-stage training approach for cross platform vehicle re-identification. When evaluated with the proposed dataset with two-stage training approach, the developed framework achieves an overall mAP of 29.76%, which is significantly higher than existing re-identification methods.

Cross platform vehicle re-identification is very challenging. Vehicles observed by two contrasting surveillance systems exhibit severe appearance, scale, and perspective changes. To estimate a match for a vehicle observed by CCTVs and UAV, a common overlap of vehicle information would be beneficial. In the dataset, there are several instances when vehicles viewed by surveillance cameras do not share common features, making re-identification more difficult. The study can be further explored by considering multiple aerial devices and CCTV cameras to collect rich information about vehicles for re-identification.

CRediT authorship contribution statement

Ashutosh Holla B.: Data curation, Methodology, Validation, Visualization, Writing – original draft. **Manohara Pai M.M.:** Conceptualization, Formal analysis, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Ujjwal Verma:** Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing – review & editing. **Radhika M. Pai:** Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] R. Makrigiorgis, N. Hadjittoouli, C. Kyrkou, T. Theocharides, AirCamRTM: Enhancing vehicle detection for efficient aerial camera-based road traffic monitoring, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2022, pp. 2119–2128.

[2] Z. Liu, W. Zhang, X. Gao, H. Meng, X. Tan, X. Zhu, Z. Xue, X. Ye, H. Zhang, S. Wen, et al., Robust movement-specific vehicle counting at crowded intersections, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 614–615.

[3] W. Liu, N. Durasov, P. Fua, Leveraging self-supervision for cross-domain crowd counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 5341–5352.

[4] X. Hao, S. Zhao, M. Ye, J. Shen, Cross-modality person re-identification via modality confusion and center aggregation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 16403–16412.

[5] X. Chen, X. Liu, W. Liu, X.-P. Zhang, Y. Zhang, T. Mei, Explainable person re-identification with attribute-guided metric distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 11813–11822.

[6] G. Wang, R. Gu, Z. Liu, W. Hu, M. Song, J.-N. Hwang, Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 9876–9886.

[7] J. Zhao, Y. Zhao, J. Li, K. Yan, Y. Tian, Heterogeneous relational complement for vehicle re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 205–214.

[8] Y. Li, J. Wu, X. Bai, X. Yang, X. Tan, G. Li, S. Wen, H. Zhang, E. Ding, Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 586–587.

[9] L. Shine, J. CV, et al., Fractional data distillation model for anomaly detection in traffic videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 606–607.

[10] N. Peri, P. Khorramshahi, S.S. Rambhatla, V. Shenoy, S. Rawat, J.-C. Chen, R. Chellappa, Towards real-time systems for vehicle re-identification, multi-camera tracking, and anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 622–623.

[11] A. Shankar, A. Poojary, V. Kollerathu, C. Yeshwanth, S. Reddy, V. Sudhakaran, Comparative study on various losses for vehicle Re-identification, in: CVPR Workshops, Vol. 2, 2019, p. 6.

[12] S. Teng, S. Zhang, Q. Huang, N. Sebe, Viewpoint and scale consistency reinforcement for UAV vehicle re-identification, Int. J. Comput. Vis. 129 (3) (2021) 719–735.

[13] H. Li, C. Li, X. Zhu, A. Zheng, B. Luo, Multi-spectral vehicle re-identification: A challenge, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11345–11353.

[14] A. Zheng, Z. Wang, Z. Chen, C. Li, J. Tang, Robust multi-modality person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3529–3537.

[15] Z. Wang, C. Li, A. Zheng, R. He, J. Tang, Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2633–2641.

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

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

[18] X. Liu, W. Liu, H. Ma, H. Fu, Large-scale vehicle re-identification in urban surveillance videos, in: 2016 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2016, pp. 1–6.

[19] X. Liu, W. Liu, T. Mei, H. Ma, A deep learning-based approach to progressive vehicle re-identification for urban surveillance, in: European Conference on Computer Vision, Springer, 2016, pp. 869–884.

[20] Y. Lou, Y. Bai, J. Liu, S. Wang, L. Duan, Veri-wild: A large dataset and a new method for vehicle re-identification in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3235–3243.

[21] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, Y. Zhang, Vehicle re-identification in aerial imagery: Dataset and approach, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 460–469.

[22] B.A. Holla, M.M. Pai, U. Verma, R.M. Pai, Vehicle re-identification in smart city transportation using hybrid surveillance systems, in: TENCON 2021-2021 IEEE Region 10 Conference, TENCON, IEEE, 2021, pp. 335–340.

[23] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 480–496.

[24] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Omni-scale feature learning for person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3702–3712.

[25] H. Chen, B. Lagadec, F. Bremond, Ice: Inter-instance contrastive encoding for unsupervised person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14960–14969.

[26] Y. Cho, W.J. Kim, S. Hong, S.-E. Yoon, Part-based pseudo label refinement for unsupervised person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7308–7318.

[27] H. Wang, J. Shen, Y. Liu, Y. Gao, E. Gavves, Nformer: Robust person re-identification with neighbor transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7297–7307.

[28] H. Ni, Y. Li, L. Gao, H.T. Shen, J. Song, Part-aware transformer for generalizable person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11280–11289.

[29] C.-W. Wu, C.-T. Liu, C.-E. Chiang, W.-C. Tu, S.-Y. Chien, Vehicle re-identification with the space-time prior, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 121–128.

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

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

[32] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15013–15022.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[35] X. Zhou, Y. Zhong, Z. Cheng, F. Liang, L. Ma, Adaptive sparse pairwise loss for object re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19691–19701.

[36] P. Antonio Marin-Reyes, A. Palazzi, L. Bergamini, S. Calderara, J. Lorenzo-Navarro, R. Cucchiara, Unsupervised vehicle re-identification using triplet networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 166–171.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[39] R. Kuma, E. Weill, F. Aghdasi, P. Sriram, Vehicle re-identification: an efficient baseline using triplet embedding, in: 2019 International Joint Conference on Neural Networks, IJCNN, IEEE, 2019, pp. 1–9.

[40] K.-T. Nguyen, T.-H. Hoang, M.-T. Tran, T.-N. Le, N.-M. Bui, T.-L. Do, V.-K. Vo-Ho, Q.-A. Luong, M.-K. Tran, T.-A. Nguyen, et al., Vehicle re-identification with learned representation and spatial verification and abnormality detection with multi-adaptive vehicle detectors for traffic video analysis, in: CVPR Workshops, 2019, pp. 363–372.

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

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

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

[44] S. Teng, S. Zhang, Q. Huang, N. Sebe, Multi-view spatial attention embedding for vehicle re-identification, IEEE Trans. Circuits Syst. Video Technol. 31 (2) (2020) 816–827.

[45] M. Naphade, Z. Tang, M.-C. Chang, D.C. Anastasiu, A. Sharma, R. Chellappa, S. Wang, P. Chakraborty, T. Huang, J.-N. Hwang, et al., The 2019 AI city challenge, in: CVPR Workshops, Vol. 8, 2019, p. 2.

[46] M. Naphade, S. Wang, D.C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, Y. Yao, L. Zheng, P. Chakraborty, C.E. Lopez, et al., The 5th ai city challenge, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4263–4273.

[47] P. Huang, R. Huang, J. Huang, R. Yangchen, Z. He, X. Li, J. Chen, Deep feature fusion with multiple granularity for vehicle re-identification, in: CVPR Workshops, 2019, pp. 80–88.

[48] A. Kanaci, M. Li, S. Gong, G. Rajamanoharan, Multi-task mutual learning for vehicle re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 62–70.

[49] T.-S. Chen, M.-Y. Lee, C.-T. Liu, S.-Y. Chien, Viewpoint-aware channel-wise attentive network for vehicle re-identification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2020, pp. 2448–2455, http://dx.doi.org/10.1109/CVPRW50498.2020.00295.

[50] S. He, H. Luo, W. Chen, M. Zhang, Y. Zhang, F. Wang, H. Li, W. Jiang, Multi-domain learning and identity mining for vehicle re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 582–583.

[51] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.

[52] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2020) 4037–4058.

[53] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[54] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[55] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.

[56] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.

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

[58] M. Li, X. Huang, Z. Zhang, Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 194–204.

[59] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, J.-N. Hwang, Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8797–8806.

[60] P. Khorramshahi, V. Shenoy, R. Chellappa, Scalable vehicle re-identification via self-supervision, 2022, arXiv preprint arXiv:2205.07613.

[61] J. Peng, H. Wang, F. Xu, X. Fu, Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification, Neurocomputing 401 (2020) 133–144.

[62] Z. Zhou, Y. Li, J. Li, K. Yu, G. Kou, M. Wang, B.B. Gupta, Gan-siamese network for cross-domain vehicle re-identification in intelligent transport systems, IEEE Trans. Netw. Sci. Eng. (2022).

[63] Y. Wang, J. Peng, H. Wang, M. Wang, Progressive learning with multi-scale attention network for cross-domain vehicle re-identification, Sci. China Inf. Sci. 65 (6) (2022) 160103.

[64] J. Peng, H. Wang, T. Zhao, X. Fu, Cross domain knowledge transfer for unsupervised vehicle re-identification, in: 2019 IEEE International Conference on Multimedia & Expo Workshops, ICMEW, IEEE, 2019, pp. 453–458.

[65] S. Girisha, M.M.M. Pai, U. Verma, R.M. Pai, Performance analysis of semantic segmentation algorithms for finely annotated new UAV aerial video dataset (ManipalUAVid), IEEE Access 7 (2019) 136239–136253.

[66] S. Girisha, U. Verma, M.M. Manohara Pai, R.M. Pai, UVid-net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 4115–4127.

[67] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.

[68] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, A. Zisserman, With a little help from my friends: Nearest-neighbor contrastive learning of visual representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9588–9597.

[69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[70] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 27 (2014).

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

[72] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[73] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

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, Nitte, India. He is currently pursuing the Ph.D. degree at Manipal Institute of Technology. His areas of interest are object detection, re-identification and deep learning for computer vision.

Dr. **Manohara Pai M.M.** (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 31 years in Teaching/Research. He holds nine patents to his credit and has published 145 papers in national and international journals/conference proceedings. 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 is also a Life Member of ISTE and a Life Member of Systems Society of India. He is also a principal investigator for multiple industry/government research projects. He has been an Executive Committee Member of the IEEE Bangalore Section, Mangalore Subsection, and the past Chair of the IEEE Mangalore Subsection. He has recently received the National Technical Teacher Award (NTTA 2022) from Ministry of Education, Government of India.

Dr. **Ujjwal Verma** received his Ph.D. from Télécom ParisTech, University of Paris-Saclay, Paris, France, in Image Analysis and his M.S. (Research) from IMT Atlantique (France) in Signal and Image Processing. Dr. Verma is currently an Associate Professor and Head of the Department of Electronics and Communication Engineering at Manipal Institute of Technology, Bengaluru, India. His research interests include Computer Vision and Machine Learning, focusing on variational methods in image segmentation, deep learning methods for scene understanding, and semantic segmentation of aerial images. He is a recipient of the ''ISCA Young Scientist Award 2017–18'' by the Indian Science Congress Association (ISCA), a professional body under the Department of Science and Technology, Government of India. Dr. Verma is the Co-Lead for the Working Group on Machine/Deep Learning for Image Analysis (WG-MIA) of the Image Analysis and Data Fusion Technical Committee (IADF TC) of the IEEE Geoscience and Remote Sensing Society. He is Guest Editor for Special Stream in IEEE Geoscience and Remote Sensing Letters and a reviewer for several journals (IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, IEEE Geoscience and Remote Sensing Letters). He is also a Sectional Recorder for the ICT Section of the Indian Science Congress Association for 2020-24. Dr. Verma is a Life Member of the Indian Science Congress Association.

Dr. **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 31 years. She has published 93 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 an Executive Committee Member of the IEEE Mangalore Subsection. She was a recipient of National Doctoral Fellowship from AICTE, Government of India.