IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 January 2023, accepted 20 February 2023, date of publication 9 March 2023, date of current version 28 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3254893

RESEARCH ARTICLE

Enhanced Vehicle Re-Identification for Smart City Applications Using Zone Specific Surveillance

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

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

Corresponding authors: M. M. Manohara Pai (mm.pai@manipal.edu), Ujjwal Verma (ujjwal.verma@manipal.edu), and Radhika M. Pai (radhika.pai@manipal.edu)

ABSTRACT Vehicle re-identification is an important feature of an intelligent transportation system as part of a smart city application. Vehicle re-identification aims at matching vehicles from images acquired by surveillance cameras at different locations. During rush hours, vehicles are densely occupied across regions such as entry/exit of gated campuses, railways, airports, educational institutions, etc. Due to this uneven flow of traffic, there is a possibility of violation of traffic rules by the vehicles that lead to a security breach. In such scenarios to speed up the re-identification process, it is justified to look into a specific group of surveillance cameras to detect and re-identify vehicles on day to day basis in near real-time. However, the existing vehicle re-identification datasets do not contain zone specific information and therefore can not be used to evaluate the performance of re-identification algorithms in different zones. In the proposed work for re-identification, a framework is developed that performs vehicle re-identification across a group of cameras that monitors varying traffic movements over an area. These areas defined as ''strategic zones'' comprise a subset of non-overlapping cameras that are installed to monitor non-uniform vehicle movements. The re-identification framework is evaluated on a novel dataset developed to understand the performance of vehicle re-identification across strategic zones. The dataset consists of videos of vehicles captured through 20 CCTV surveillance cameras that are grouped into four different zones. Various experiments are conducted to study the performance of re-identification across four zones using a deep neural network with triplet loss, L2 regularization, and re-ranking. The experiments conducted with an image dimension of 224×224 have demonstrated an overall mAP of 77.22%. Also, for each of the four zones a mAP of 82.16%, 69.1%, 66.5%, and 75.76% is achieved. The experimental results demonstrate huge variations in the accuracy of vehicle re-identification method across different zones. Therefore, the study assess the possible measures that can be taken to improve the performance in individual zones for an accurate vehicle re-identification in intelligent transport system.

INDEX TERMS Keyframes, regularization, surveillance, triplet loss, vehicle re-identification.

I. INTRODUCTION

A great deal of attention has been paid to surveillance in public areas where pedestrians and vehicles are more prevalent. Providing robust security measures is a primary concern for any government and supporting organization due to the increasing cases of thefts, vandalism, etc. To provide robust security, remote monitoring systems, mainly CCTV cameras, are widely used. Applications of surveillance systems include smart transportation, monitoring, controlling traffic signals

The associate editor coordinating the review of this manuscript and approving it for publication was Haixia Cui.

and traffic movements [1], agriculture [2], and wildlife monitoring [3].

In recent years, a trend has been followed in major cities to develop it as ''smart cities''. The motive of the smart city is to strengthen economic growth, improve the quality of living for people, and harness technology that yields better outcomes. Among the initiatives taken under the smart city development, providing a robust infrastructure for Intelligent Transportation Systems (ITS) in traffic surveillance is still a primary concern. ITS aims to relieve major problems that frequently affect road transportation, such as minimizing the congestion to ease the traffic flow, avoiding potential accidents that impact the environment, and improving the air quality by reducing the travel delay. It also guides the travelers with preliminary information about traffic and real-time traffic density information that enables the user to minimize their daily travel time. ITS processes the information acquired by surveillance cameras to automate traffic monitoring, thereby guiding the traffic authorities to take advert possible mishaps caused by commuters. The acquired surveillance footage is useful in ITS to carry out vehicle detection [4], [5], vehicle counting [6], vehicle re-identification [1] and tracking [7], [8]. Currently, more efforts are put in place in computer vision to develop a robust vehicle re-identification framework. Vehicle re-identification aims to match and retrieve an identical vehicle that is appeared across a network of surveillance cameras [9]. Using surveillance camera information, the presence of a vehicle observed at a particular camera is queried across all surveillance cameras. Vehicle re-identification is challenging due to variation in viewpoint, illumination, scale, etc. Besides, vehicles of the same model and color pose a tough task for re-identification.

For traffic surveillance, a network of surveillance cameras is usually installed across accidental prone areas such as highways, freeways, intersection/junction points, etc. It is also commonly seen around the entry and exit of gated campuses such as railways, airports, educational institutions, etc., to monitor the traffic daily. However, the distribution of vehicles around these areas need not be uniform. During rush hours, more vehicles will be along the significant intersections and junction points compared to an arterial road. Additionally, due to the occurrence of unanticipated events, dense occupancy of vehicles accumulated at arterial regions could result in a roadblock. In the context of ITS, deploying ample surveillance cameras to monitor every possible event is costly. However, an adequate number of surveillance cameras must be put in place at precise locations to gather rich information on traffic flow. The information gathered by these surveillance cameras should be sufficient enough for ITS to make proper future decisions that are unanticipated.

The existing re-identification methods do not dissect individual regions under surveillance; rather, they utilize the *entire set* of surveillance cameras to re-identify the vehicles. For an area monitored by a group of cameras that are regularly exposed to traffic breaches, the global approach to re-identify vehicles fails to understand the impact of re-identification models. Therefore, there is a need to analyze the performance of re-identification algorithms for individual regions. In this study, the entire surveillance area is divided into regions that are referred to as zones. These zones comprise a subset of non-overlapping cameras installed within a few meters of separation as commonly observed at intersection/junction points, entry/exit of a gated campus, etc. This study focuses on studying the performance of vehicle re-identification algorithms in each zone rather than the entire surveillance area. This approach enables ITS to focus only on a group of surveillance cameras where there is a breach of traffic due to unforeseen events and subsequently detect and re-identify those vehicles in individual zones. In such scenarios, tracking these vehicles across different zone specific cameras is applicable if the information about vehicles traveling in these zones is known in advance. Specifically, this study attempts to answer this question: *Is the performance of vehicle re-identification algorithm dependent on the location/placement of the surveillance cameras?*. This finer zone-wise analysis would provide useful insights into the performance of the vehicle re-identification algorithms and any corrective measures to further improve the re-identification accuracy. Further, zones with high traffic density/traffic breach can be targeted as strategic zones. In ITS for these strategic zones, an utmost priority can be given to providing a robust traffic measure for a smoother movement of vehicles. Though the datasets such as VehicleID [10], VeRi-776 [11], VeRi-Wild [12] contains vehicle information observed across a network of surveillance cameras, they utilize *entire* surveillance cameras to supervise the task of re-identification. To the best of our knowledge, these datasets lack the zone specific vehicle information that is useful for re-identification across a group of cameras. A first-of-its-kind study is conducted to facilitate such scenarios by developing a re-identification framework to re-identify the vehicles across different zones.

In this study, videos acquired from CCTV surveillance cameras of an educational institute are analyzed for zone specific vehicle re-identification. For this study, the entire educational institute is divided into four zones. For a given query vehicle image, the features are computed from the re-identification network and matched with the features of the gallery images from the *same zone* as the query image (Figure 1). The major contributions of this paper can be summarized as follows:

- A novel zone specific vehicle re-identification dataset is developed for performing vehicle re-identification. Dataset consists of 81 vehicle identities observed at 20 different surveillance cameras. 2,300 manual annotations are provided for vehicle identities observed in 37,722 keyframes.
- A zone specific vehicle re-identification framework is proposed for re-identifying the vehicles across four zones. A standard CNN ResNet50 [13] architecture is used to train the re-identification network with triplet loss and L2 regularisation.

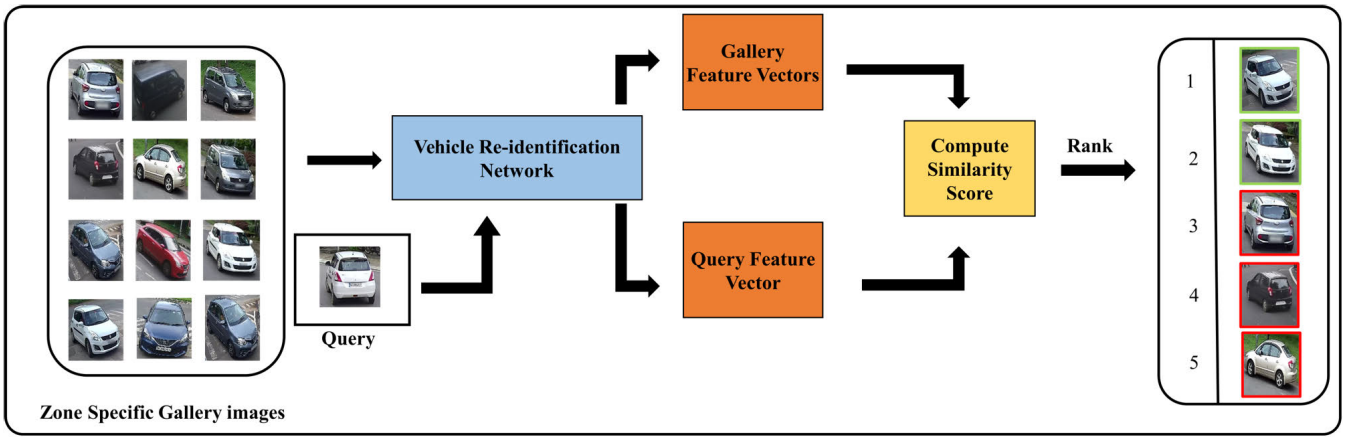

FIGURE 1. Outline of zone specific vehicle re-identification for CCTV videos. For a given vehicle query during inference, its presence is determined by extracting the features of the query vehicle with those vehicles observed across individual zones. Finally, the gallery images are ranked with the vehicles resembling the query at the top.

- An ablation study is conducted to evaluate the performance of vehicle re-identification across the zones for different sets of parameters.

II. RELATED WORK

Computer Vision-based approach has been used to analyse images/videos for various applications such as remote sensing [14], [15] precision agriculture [16], [17], surveillance [18], [19], [20], bio-medical imaging [21], [22], etc. In the past few years, several works are contributed to conduct vehicle re-identification using surveillance cameras. In the present study, a zone-specific vehicle re-identification framework is designed to re-identify vehicles that appear in subsets of surveillance cameras in different zones. The developed framework utilizes the detected instances of vehicles using a standard CNN object detector. There have been several developments in developing vehicle detectors that are used along with vehicle re-identification and tracking. Since the present study focuses on vehicle re-identification, a summary of vehicle detection methods is provided that aid in performing re-identification. The following paragraphs summarise recent developments in vehicle re-identification that utilize the bounding boxes of vehicles generated by state-of-the-art object detector models.

A. VEHICLE DETECTION FOR RE-IDENTIFICATION

Object detection, in the traditional approach, involves selecting the regions of interest (vehicles in the present study) and extracting the visual and semantic features from HOG [23], SIFT [24], and Haar-like features [25] and then classifying these features using Support Vector Machine, Deformable Part-Based Models, and AdaBoost Classifiers, etc. Notable contributions are contributed to detecting vehicles using these traditional approaches [26], [27], [28].

With the advancement of convolutional neural networks, many object detector networks have been developed to address vision applications. The object detectors fall into two categories: region proposal based methods and regression/classification based methods [4]. The methods based on region proposals are divided into two phases. Region proposals are generated using different strategies [4] and these regions are utilized to classify and localize objects using CNN extractors. Prominent detectors that fall under these categories are R-CNN [29], Fast-RCNN [30], Faster-RCNN [31] and Mask-RCNN [32]. R-CNN and Fast-RCNN methods are computationally expensive since each region proposal has to be predicted to determine if an object is present. Moreover, the Mask-RCNN approach contains a classification and regression branch and additional branches to predict the segmentation mask. Some of the recent articles made use of Faster-RCNN and Mask-RCNN that dealt with triggering vehicle actions for autonomous vehicles [33], to generate segmentation masks for vehicles for multi-camera vehicle re-identification and tracking [34], [35].

Regression/classification-based object detectors are one-step object detectors that map the image pixel information directly to predict the bounding box coordinates and class probabilities, thereby reducing a considerable amount of time compared to region proposal methods. Notable object detectors such as variants of YOLO and SSD. YOLO was introduced by Redmon et al. [36] contains an end-to-end framework that comprises feature extraction which aids in predicting the confidence scores for different class categories and bounding box coordinates. Authors proposed YOLOv2 in [37] that adopts strategies such as batch normalization, dimension clustering, anchor boxes, and multi-scale training. Authors developed Single-Shot Detectors in [38] which add feature extractor layers at the end of the backbone network to determine the offsets of bounding boxes for different scales and aspect ratios. YOLOv3 [39] uses a deeper feature extractor network, Darknet-53 that performs predictions at three different scales. It outputs the bounding box coordinates of the object with a class confidence score. Researchers conducted a study in [40] that aimed to detect vehicles observed by UAV's. They compared the vehicle detection performance of both YOLOv3 and Faster-RCNN.

Authors concluded their study by stating that YOLOv3 outperforms Faster-RCNN by faster execution and processing time. In [41], authors developed a modified implementation of YOLOv3 to address the vehicle bounding box localization uncertainty for autonomous driving. To improve speed and accuracy in detecting objects, YOLOv4 [42] introduced Spatial Pyramid Pooling (SPP), which chooses the most important contextual information. CSPDarkNet53 is used as a backbone feature extractor network along with Path Aggregation Network for detecting objects at different levels. The Bag-of-Freebies (BoF) and Bag-of-Specials (BoS) strategies are introduced to improve object detection. Authors in [43] made use of the YOLOv4 object detector to identify vehicles for re-identification and tracking across surveillance cameras. Authors in [44] modified the YOLOv4 architecture mainly to detect target objects in remote sensing images. For aerial images present in the DOTA dataset, their modified framework aims to detect four different types of targets. To improve detection accuracy for each target object, a Non-maximum Suppression (NMS) threshold with varying values for each category is chosen. Additionally, rather than having a unified anchor box dimension at various scales, they proposed dividing the targets into three distinct scales. This was done using an anchor frame distribution that applies K-means clustering to group these into three distinct scales.

YOLOv5 [45] has similar components to YOLOv4 with CSP-DarkNet53 as the backbone feature extractor network and SPP and PANet. A focus layer is integrated by replacing the first three layers of the backbone network to improve mean Average Precision (mAP). Different variants of YOLOv5 are available that have minor differences in the number of layers used in the network architecture. Some notable contributions toward vehicle detections are presented in [46] and [47]. YOLOv6 [48] is primarily designed for industrial applications. YOLOv6 uses a hardware-friendly backbone network, and decoupled head architecture for calculating regression, objectness, and class confidence scores. The architecture uses VariFocal loss as classification loss with SIoU or GIoU loss as regression loss. YOLOv7 is the most recent variant of the YOLO object detector family. YOLOv7 [49] is developed based on modifying the Efficient Layer Aggregation Network (ELAN) named E-ELAN (Extended-Efficient Layer Aggregation Network). The modified architecture aims to increase detection performance at inference time without affecting the gradient flow paths of the network. YOLOv6 and YOLOv7 are presented recently as an object detector framework with a contribution limited to a study presented in [50] for vehicle re-identification and tracking.

Transformers in computer vision have grown significantly in recent years, addressing problems related to object classification, detection, segmentation, etc. Frameworks that rely on transformers learn the features of objects by computing multi-head attention scores that determine the weightage of patches over neighboring patches. Some of the notable transformer based object detectors are ViT [51], Swin [52], DINO [53], DETR [54] etc. However, these transformer based object detectors are data intensive for effective learning along with larger inference time than compared to CNN based object detectors for performing predictions.

B. VEHICLE RE-IDENTIFICATION WITH SUPPLEMENTARY INFORMATION

Authors in [55] developed a vehicle tracking framework, namely constrained multiple-kernel (CMK) tracking, to address the scenarios of a vehicle subjected to occlusion. The developed framework uses a vehicle localization approach by modeling multiple kernels and associating these with different forms of 3D deformable vehicle models along with camera calibration. They evaluated their framework on the NVIDIA AI CITY Challenge dataset. To detect vehicle instances, they combined the predictions generated by YOLO9000 [37] and SSD [38] object detectors. They estimate camera calibration by using segmentation to generate foreground blobs of humans appearing in frames. For tracking vehicles, their framework combines the predictions of the Kalman-filtering framework with CMK. Their approach of utilizing 3D models also aids in localizing the license plates of vehicles to perform re-identification. The authors extended the above work by presenting a vehicle tracking and recognition method to track and recognize vehicles across cameras [56]. To re-identify the vehicle, the method utilizes a 3-D deformable vehicle model to extract vehicle attributes such as license plate information, type of vehicle, and color. In cases where the resolution of the license plate is good enough, the license plate is segmented and recognized using OCR-based License Plate Recognition (LPR). For vehicle recognition, they have used SSD to detect the license plate and estimate the similarity of the detected plate across frames. Re-identification of vehicles may be enhanced by using license plate information. In any case, every license plate is regarded as uniquely identifiable information that may breach the privacy of the user [57]. In certain multilingual countries, despite strict government regulations, some drivers misuse the norms and have license plates with different characters. Therefore, it will be challenging to scale and adapt LPR to learn non-uniform license plate information. While vehicle attributes may aid in better re-identification, the trade-off may be seen in providing/modeling this additional information with annotations and labeling, which requires considerable effort and time.

In [58], the authors addressed vehicle re-identification problems such as data annotation difficulties and visual appearance mismatch across two vehicle re-identification datasets. Here the authors have proposed an adaptive feature learning method for vehicle instances observed across different datasets. To re-identify the vehicles for datasets that do not contain ground truth labels, authors made use of available vehicle re-identification datasets such as VeRi [59], CompCars Surveillance [60], and BoxCars [61] which contain ground truth information such as identity of a car, model, and other supplement information. A CNN model is further trained using this information to learn vehicle features

to improve the re-identification of vehicles. To generate pairs of vehicles that convey a positive match, those vehicles appearing in successive frames are considered, while with the fact that the exact vehicle cannot appear at multiple locations, vehicle detection pairs for a given frame at a specific timestamp are considered as negative pairs. Their method fails in re-identifying the vehicles containing discriminative information. Authors in [62], incorporated the discriminative local cues and global information of the vehicles to perform multi-scale attention for re-identifying the vehicles. A bilinear interpolation is used to obtain features at different scales fused with feature maps generated by attention blocks. These blocks are utilized to mine discriminative features of vehicle.

A dual attention re-identification network is proposed by authors in [63] that selectively scores the vehicle parts with higher attention scores. The framework extracts the vehicle features by using a dual branch CNN network. This network makes use of self-attention layers that learns to identify the different region of interest in vehicle. Sub-networks later refine regional features of the ROIs to obtain more coarse/fine level vehicle recognition. Recently several works to perform vision-related tasks such as image recognition detection and segmentation using the transformer network have been contributed. A CNN and transformer-based vehicle feature fusion is performed to re-identify the vehicles [64]. ResNetmid learns semantic and global features, and a Swin transformer [52] is used to extract the vehicle features at different scales. Vehicle features are extracted and fused at inference to rank the gallery vehicle images for a given query. The authors in [65] compared vehicle re-identification models for trucks utilizing axle spacing and vehicle lengths. They formulated the re-identification as a mathematical assignment problem where models such as Naïve Bayes, Bayesian models, and Gaussian mixture models.

To address a generalization problem associated with cross-dataset vehicle re-identification, authors in [66] presented a framework that fuses both visual and spatio-temporal information. The authors highlighted the drawbacks when a re-identification network trained on a single dataset may fail to generalize the representation of vehicles for different vehicle re-identification datasets. Their framework comprises a siamese network that consists of parallel networks of shared CNN feature extractor layers. Data augmentation is performed on vehicles appearing in the source dataset. These networks aim to learn vehicle features from the source dataset and the augmented version of the vehicle from the same dataset using a siamese classifier. Furthermore, for a pair of images appearing in an unlabeled dataset, knowledge about spatio-temporal features of vehicles is learned using transfer learning techniques from a siamese network.

Vehicle re-identification and abnormality detection are performed in [67]. Their approach consists of three modules: a deep metric embedding module, vehicle attribute extraction module, and re-ranking. Authors utilized the information of vehicles such as vehicle category, vehicle type, and pose information to train a classifier for performing re-identification. They have estimated the pose of the vehicles to determine the similarity of a query vehicle image across the collection of vehicles observed in the gallery. As a part of optimization, authors have used the bag-of-words approach to re-rank the matched gallery-query images.

Authors in [68] have proposed a framework to perform vehicle re-identification using vehicle location and time stamps. They have addressed how re-identification is challenging to achieve by considering just image information. A feature ensemble technique to increase discriminative information of the vehicle, a system to extract the location and time information by considering raw video input, and a time stamps to remove incorrect images during re-ranking were developed. As a part of extracting feature representations, they have used CNN-based DenseNet-121, which is trained using triplet and cross-entropy loss. The network has been trained under three scenarios to generate a discriminative representation of the vehicle. A temporal pooling is performed on the gallery images to obtain temporal information. The query and the gallery images are matched with location and time stamps information to perform re-identification. The camera locations are used to infer the presence of a vehicle in the scene. The number of gallery images considered has been restricted by constructing a transfer matrix. This matrix provides information about the distance between cameras installed to perform re-identification. Using this distance matrix, the maximum time required for a vehicle that appears in a pair of cameras can be determined, thereby eliminating those images that exceed the threshold distance. The matrix refines the gallery images by eliminating redundant images that do not contribute toward vehicle re-identification.

As a part of the AI City challenge [69], [70] several works were contributed by considering the subset of cameras to track and identify anomalous activities observed at different areas where CCTV cameras are installed. In [71], authors developed a locality-aware multi-camera tracking algorithm. Here the vehicles are initially detected, and reliable tracklets are computed from initial detections. For each scene that contains a crossing or a turn, the trajectories are linked to perform cross-camera re-identification and tracking. The vehicles in the scenes are selected using ROI refinement, spatial-temporal smoothness, and scene division technique. In [72], as a part of the AI City challenge, authors have addressed how to identify anomalous activities such as identifying stalled vehicles in a scene. Here a road mask is created to identify the vehicles on the street. From the known fact that the movement of anomalous vehicles compared to other slower vehicles, these vehicles are flagged with fewer movements than others. A pre-trained Mask-RCNN [32] is used to generate a bounding box for vehicles appearing in the scene. Stalled vehicles are then identified where there is an increase in overlap of a bounding box for vehicles appearing in successive frames.

Apart from the above works for re-identifying vehicles in CCTV cameras, studies specific to re-identifying

vehicles using UAVs have been performed in recent years [73], [74]. Along with these works, a vehicle re-identification framework is presented in [75] to re-identify the vehicles observed across two different modalities(CCTVs and UAV). However, the current study focuses on the impact of vehicle re-identification across zones monitored by surveillance cameras.

The approaches mentioned above dealt with issues of improving re-identification performance. In many cases, re-identification methods considered complete surveillance camera information to address different use cases. Nevertheless, there is a lack of study that performs vehicle re-identification over a given area with a subset of cameras. Performing zone specific re-identification enables us to analyze which area is heavily exposed to traffic mishaps, abnormal events, violations of traffic rules, etc. In doing so, stringent measures can be taken in the future in those areas that are subjected to severe traffic outbreaks.

Note that most existing works on available datasets such as Comprehensive Cars [60], VehicleID [10], VeRi [59] etc., the vehicle information is queried across *all* surveillance cameras to estimate the presence of a vehicle. To the best of our knowledge, no existing dataset studies the performance of vehicle re-identification algorithms for individual zones.

III. METHODOLOGY

The surveillance data is acquired from an educational institution to study the performance of vehicle re-identification algorithms. The details of surveillance data used for re-identification are discussed in Section III-A. As a preprocessing step, a shot boundary detection (Section III-B) is applied to generate keyframes. For the detected identical vehicle instances using a standard object detector, the license plate is blurred and assigned a unique vehicle identification number. Subset of identical vehicles are used for performing vehicle re-identification (Section III-C). During the inference, the presence of the identical vehicle given as a query is looked upon as the vehicle appeared in each zone (Section III-D).

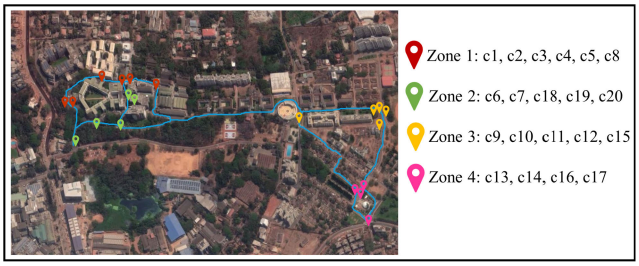

FIGURE 2. Outline of a map with zone specific CCTV camera locations.

A. ZONE SPECIFIC VEHICLE RE-IDENTIFICATION DATASET

The videos for this study have been acquired from the surveillance cameras (CCTV) installed on the campus of our educational institute (Manipal Institute of Technology, Manipal, India). Of the entire cameras available on the campus (Total area: 188 acres), the cameras considered for this study are taken such that the probability of traffic movements is non-uniform. Camera locations include entry/exit of campus, academic section, hostel premises, etc. The zones are manually identified for each of the chosen 20 cameras, and the subset of cameras is grouped in each zone. On average, each zone comprises 4 to 6 cameras. Figure 2 shows the camera identities corresponding to each zone. Zone 1 and 2 are comprised of surveillance cameras that are located in the academic area within the campus, where dense traffic movements are observed very often. Zone 3 consists of a fleet of cameras situated near roads that connect different hostel blocks, and zone 4 contains group surveillance cameras at the entry/exit point near the hostel premises. The data is gathered from Hikvision surveillance cameras with a resolution of 1920 × 1080p at 20fps. Table 1 summarizes the CCTV surveillance data information gathered for two days. Also, Figure 3 illustrates the sample images from surveillance cameras particular to each zone.

During the inference, vehicle re-identification is evaluated for each zone individually. Two days of videos regarding traffic movements are collected using surveillance cameras. 81 identical vehicles were identified across 20 cameras for two days of collected data. During data acquisition, there were dense traffic movements across zones 1, 2, and 4 compared to zone 3. Using these observations, during inference, for every probe vehicle image the performance of the re-identification framework is analyzed for non-uniform vehicle movements across zones.

TABLE 1. Description of data acquisition for CCTV videos.

Total CCTV cameras	20 Cameras
Frame resolution	1920×1080p
Frame rate	20 fps
Duration of a CCTV data	15 to 40 min

B. DATA PREPROCESSING: SHOT BOUNDARY DETECTION

Vehicle re-identification involves identifying a vehicle of interest in the respective installed surveillance camera and thereby querying the identified vehicle to check for its presence if other surveillance cameras observe it. Table 1 shows that the data acquired by CCTV surveillance cameras are at 20 fps. Hence the change of information from frame to frame is minimal.

Processing every frame for performing vehicle re-identification is expensive, and it can be minimized by considering specific frames called keyframes using a shot boundary detection algorithm. Moreover, the duration of each video ranges from 15 to 40 min. Hence processing every frame with its original frame dimension for shot boundary detection is computationally expensive. Initially, every frame is resized to a dimension of 512 × 512. The frames are further divided into a non-overlapping grid of 16 × 16. For

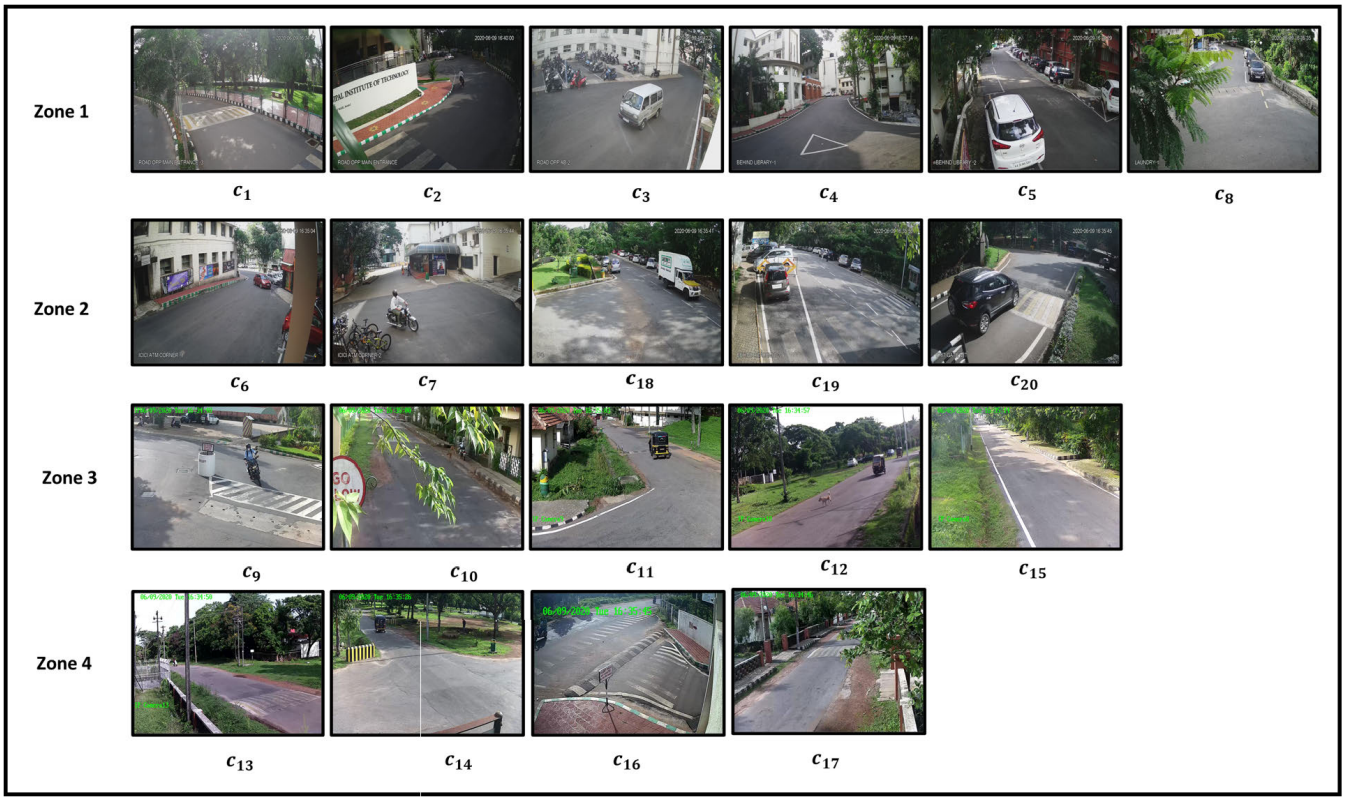

FIGURE 3. Few keyframes from the various cameras along the four zones studied in this work. The CCTV cameras are numbered as c_i.

two consecutive frames, the histogram difference is computed using the Chi-Square distance as given in the equation (1)

$$D_i = \frac{1}{N}\sum_{k=1}^{N}\frac{(H_i(I_k) - H_{i+1}(I_k))^2}{H_i(I_k)} \tag{1}$$

where $H_i(I_k)$ denotes the histogram of k^{th} image patch I_k of i^{th} frame. Similarly, $H_{i+1}(I_k)$ represents the histogram of k^{th} image patch I_k of $(i+1)^{th}$ frame, D_i represents the average histogram difference between i^{th} and $(i+1)^{th}$ consecutive image frames and N represents the total number of grids in an image. For a given input image of size 512×512, 1024 grids are obtained. The difference D_i is calculated for every pair of consecutive frames and shot boundary is identified as given in equation (2)

$$\text{Shot\_boundary} = \begin{cases} True & D_i - D_{i+1} > T_{shot} \\ False & otherwise \end{cases} \tag{2}$$

The value for T_{shot} is determined experimentally. Figure 4 shows the variation of this distance for a particular video. It can be seen that the shots can be identified by selecting $T_{shot} = 0.2$. If $D_i - D_{i+1}$ is greater than the threshold value T_{shot} a shot boundary is identified. This study identifies the middle frame within a shot as a keyframe. Later re-identification of vehicles is carried out for the identified keyframes by considering it with the original image resolution (1920×1080p).

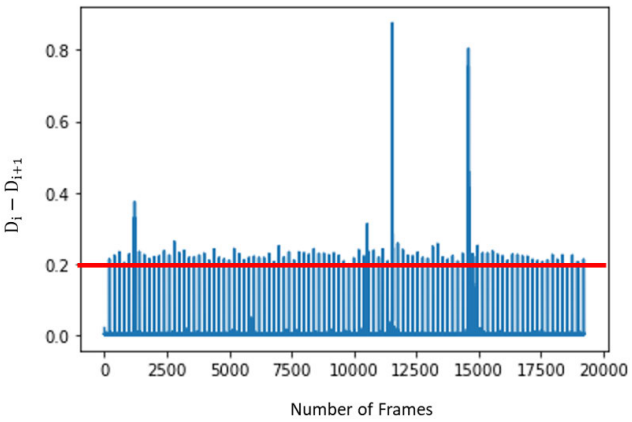

FIGURE 4. Histogram difference variation ($D_i - D_{i+1}$) for a particular CCTV video.

The generated keyframes are given as an input to a standard object detector YOLOv3 [39] to detect the presence of vehicles in these keyframes. For each detected vehicle by YOLOv3, those vehicles (identical vehicles) appearing at a minimum of two surveillance cameras are manually annotated using the Microsoft Visual Object Tagging tool (MS Vott). To maintain the privacy of the owner of the vehicles, the license plate of each vehicle observed in the selected keyframes is manually blurred. Among the total identical vehicles observed, a summary of these vehicles' appearances corresponding to the four zones is shown in Table 3.

TABLE 2. Data annotation statistics for CCTV videos.

	Number of Keyframes	Number of frames annotated	Number of Identical vehicles
Train Data	20,578	1,317	46
Test Data	17,548	983	35

TABLE 3. Details of appearance of identical vehicles across four zones.

	Zones	Appearance of Identical Vehicles	Number of Keyframes
Train	Zone 1	18	385
	Zone 2	37	662
	Zone 3	12	88
	Zone 4	12	182
Total train Images			1,317
Test	Zone 1	8	138
	Zone 2	18	341
	Zone 3	13	129
	Zone 4	21	375
Total test images			983

C. VEHICLE RE-IDENTIFICATION

An existing deep neural network is used as the backbone architecture to perform vehicle re-identification. The principle behind performing vehicle re-identification is inferred from work person re-identification [76]. The re-identification is performed with an initial minimal setup followed by a few modifications to the architecture to carry out a post-set of experiments. The workflow of each experiment undertaken to perform re-identification is summarized below.

1) MINIMIAL SETUP

The Figure 5 describes the overall architectural diagram for vehicle re-identification. To learn vehicle representations, ResNet50 [13] is considered a backbone model. The images of vehicle identities are passed through a ResNet50 architecture. A global average pooling layer is added after the fifth residual block of ResNet. This layer squeezes the spatial dimension of the feature maps. Except for the minimal setup, the post set of experiments uses a clipping layer inserted after the pooling layer, which performs an element-wise value clipping to ensure that the values are in close intervals. These values are further normalized using a BatchNormalization [77] layer. The network's last layer comprises a fully connected dense layer containing hidden units equal to the total vehicle identities observed in the training phase of vehicle re-identification. The network is trained using categorical cross-entropy loss to determine the probability of a vehicle instance belonging to the same vehicle identity. A label smoothing regularization is adopted while training the model with categorical cross-entropy loss. The architecture is initialized with pre-trained ImageNet [78] weights for the first four residual blocks of ResNet50.

A label smoothing regularization is adopted while training the model with categorical cross-entropy loss. For a given set of training samples with ground truth labels $y \in \{1, 2, \ldots .K\}$ the equivalent one-hot encoded label $y_{he}(i)$ is equal to 1 if the given index i is as the same as label y and 0 otherwise. Using the hyperparameter $\alpha \in (0, 1)$ the smoothed label

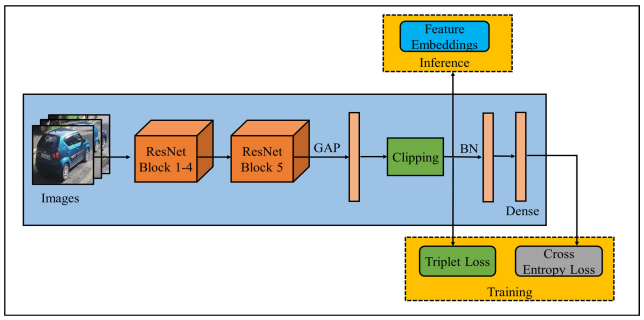

FIGURE 5. Vehicle re-identification architecture. For minimal setup experiment, the blocks colored green are excluded.

is calculated as

$$y'_{he}(i) = (1 - \alpha)y_{he}(i) + \frac{\alpha}{K} \tag{3}$$

Here α determines the amount of label smoothening, and K denotes the number of class labels. The random erasing [79] technique is adopted while training the network, which considers a rectangular portion of an image and erases its pixels with random values. It is used to handle those vehicle images subjected to partial occlusion. At the inference stage for all vehicle re-identification experiments, the vehicle feature embeddings after global average pooling are extracted as the representations, and the cosine similarity distance metric is adopted to quantify the correlation among vehicle images.

2) VEHICLE RE-IDENTIFICATION+TRIPLET LOSS

Apart from the minimal setup, the re-identification network is trained with the triplet loss [80]. As illustrated in the Figure 5, for the generated feature embeddings after applying global average pooling, the re-identification network utilizes these embeddings to compute triplet loss. Given an anchor image of a vehicle instance, triplet loss minimizes the distance between the anchor and the vehicle instance of the same identity while maximizing the distance of a vehicle instance belonging to a different identity. Triplet loss formulates the decision to classify the vehicle identity belonging to a particular class using a margin parameter.

3) VEHICLE RE-IDENTIFICATION+TRIPLET LOSS+L2 REGULARIZATION

Another set of experiments conducted for vehicle re-identification is the utilization of L2 regularization. Here the selected regularization factors are updated adaptively through backpropagation. For a set of neural network parameters $T = \{w_m \mid m = 1, \ldots M\}$, a lower regularization factor is applied to initial and higher regularization factors at deeper layers. w_m denotes either network parameters such as kernel, gamma, bias, beta and dense. The regularization factor λ_m is associated with each network parameter such that

$$L_\lambda(T) = L(T) + \sum_{m=1}^{M} (\lambda_m ||w_m||_2^2) \tag{4}$$

Here the $L(T)$ is the weighted sum of triplet loss and categorical cross entropy loss and $L_\lambda(T)$ is the revised function

after applying regularization. The regularization parameter $\lambda_m \in \mathbb{R}_+$. To ensure the regulraization factors λ_m always have positive values, a hard sigmoid function is applied. It is defined as

$$f(x) = \begin{cases} 0, & if\ x < -c \\ 1, & if\ x > c \\ \frac{x}{2c} + 0.5, & otherwise \end{cases} \quad (5)$$

During experimentation, the value for c is taken as 2.5.

Further details regarding the L2 regularization can be found in [76].

D. ZONE SPECIFIC VEHICLE RE-IDENTIFICATION

The performance of vehicle re-identification during inference is studied using the vehicle information gathered by a group of cameras particular to each zone. As pointed out, the 20 surveillance cameras are clustered into four zones. As illustrated in the Figure 1, vehicle re-identification is evaluated for each zone. Here the gallery set consists of those images of vehicle identities observed by cameras belonging to individual zones. The appearance of a query vehicle is determined by verifying its vehicle identity and camera identity with those zone specific vehicle identity images appearing in the gallery. Further, using the gallery and query embeddings, a similarity score is computed to rank the appearance of query images to zone specific vehicle images. Vehicle re-identification scores are estimated individually for each zone using the computed similarity matrix.

IV. RESULTS AND DISCUSSION

The section highlights the experimental details carried out to perform vehicle re-identification. Also, the subsequent sections give the particulars of re-identification scores obtained using the available metrics.

A. IMPLEMENTATION DETAILS OF RE-IDENTIFICATION NETWORK

The annotated vehicle instances observed on two days are considered for vehicle re-identification experiments. To train the re-identification network, 46 vehicle identities appearing across 20 surveillance cameras are considered. The experiment is carried out for two different image dimensions, 128 × 128 and 224 × 224, respectively. Accordingly, the ResNet50 architecture is implemented separately for two different image dimension vehicle identities. The network is trained using the Batch Hard [81] triplet loss variant for different batches of inputs. For each experiment, the vehicle identities, i.e., (P) [81] considered training the triplet loss network, are taken at multiples of 3 beginning from 3 to 15. Adam optimizer is used to train the network with the learning rate of 2e−4 for 200 epochs. The learning rate is divided by ten once after the performance on the validation plateaus. The selected regularization factors λ_m in Equation 4 are set to 0.005 in the experiment. A probability of 0.5 is taken to perform Random Erasing data augmentation. During inference, 983 gallery images comprising 35 vehicle identities observed at 20 surveillance cameras are taken as gallery/test set. 35 vehicle identities are considered in the inference phase to determine their existence in the gallery set. To improve the accuracy of re-identification, k-reciprocal re-ranking [82] is adopted to refine the initial query to gallery set ranking. Starting with the minimal setup for vehicle re-identification, the performance of re-identification is studied for each zone individually. Further the network is evaluated for post-set of experiments using Triplet Loss, L2 regularization, and re-ranking techniques.

TABLE 4. mAP and rank-k score for minimal setup.

Image Dimension	Mode	mAP	rank-1	rank-5	rank-10	rank-20
128 × 128	20 Cameras	40	50	83.3	83.3	83.3
	Zone 1	41.9	50	66.6	83.3	83.3
	Zone 2	29.08	25	43.7	62.5	81.25
	Zone 3	18.61	18.18	36.36	45.45	54.54
	Zone 4	33.78	55	70	80	80
224 × 224	**20 Cameras**	**59.82**	**66**	**83**	**83**	**83**
	Zone 1	**59**	**66**	**83**	**83**	**100**
	Zone 2	**41.01**	**37.5**	**56.25**	**87.5**	**93.75**
	Zone 3	**13.21**	**18.18**	**27.27**	**27.27**	**36.36**
	Zone 4	**55**	**90**	**90**	**90**	**90**

1) EVALUATION METRICS

To evaluate the performance of vehicle re-identification, available metrics such as the mean-average-precision (mAP) and rank-k accuracy are adopted. For each query, a distance matrix is computed with all the gallery images. The distance matrix measures the similarities of a query to gallery images. It is sorted such that all vehicle instances similar to the query appear at the initial entry of the matrix, and different vehicle instances appear at last. Existing vehicle re-identification datasets such as VehicleID [10], [59], VeRi-776 [11], VeRi-Wild [12] does not include zone specific information. Hence the proposed re-identification framework cannot be evaluated on these datasets. In this work, the re-identification metrics for a given set of query images are calculated for two cases 1) Entire 20 surveillance cameras 2) zone specific surveillance cameras.

1) **Entire 20 surveillance cameras:** Here the occurrence of a given query is determined across each surveillance camera. From the distance matrix estimated for each query image, the gallery samples are discarded if they have been retrieved from the same surveillance camera. As a result, more focus is laid on performing cross-camera re-identification. Accordingly, both mAP and rank-k accuracy are calculated.
2) **Zone specific surveillance cameras:** Here the appearance of a given query is determined by looking into individual zone camera information. While calculating the metrics such as mAP and rank-k scores, solely for each zone, the identity of the query vehicle is verified with vehicle images that appeared across considered zone cameras. Apart from the minimal setup

TABLE 5. mAP and rank-k score for ResNet50+Triplet loss.

Image Dimension	P	Mode	mAP	rank-1	rank-5	rank-10	rank-10
128 × 128	3	20 cameras	51.36	60.6	78.7	87.8	90.9
		Zone 1	53.1	66	83.3	83.3	100
		Zone 2	52.2	56.2	56.2	75	75
		Zone 3	22	36.36	36.36	54.54	54.54
		Zone 4	46	50	75	90	90
	6	20 cameras	52.32	69.69	78.78	90.9	93.93
		Zone 1	44.54	66.6	66.6	83.3	83.3
		Zone 2	53.26	50	62.5	81.2	81.2
		Zone 3	36.6	45.4	54.54	54.54	63.63
		Zone 4	55.4	51.1	66.6	66.6	90
	9	20 cameras	46.3	63.63	78.78	93.93	96.96
		Zone 1	49.36	50	50	66.66	83.3
		Zone 2	46.29	43.75	56.25	62.5	75
		Zone 3	20.79	18.18	36.36	45.45	63.63
		Zone 4	47.23	65	85	95	95
	12	20 cameras	42.83	57.57	69.69	84.84	93.93
		Zone 1	42.98	50	66.66	66.66	83.33
		Zone 2	49.83	43.75	62.5	81.25	81.25
		Zone 3	27.81	36.36	36.36	45.45	63.63
		Zone 4	48.7	60	80	95	95
	15	20 cameras	42.06	51.51	66.66	81.81	90.9
		Zone 1	45.89	50	50	66.66	66.66
		Zone 2	48.366	43.75	50	68.7	75
		Zone 3	13.227	18.18	27.27	36.36	63.63
		Zone 4	44.8	60	80	80	90
224 × 224	3	**20 cameras**	**62.37**	**69.69**	**84.84**	**93.9**	**93.9**
		Zone 1	**71.27**	**66.6**	**100**	**100**	**100**
		Zone 2	**65.97**	**62.5**	**81.25**	**93.7**	**93.7**
		Zone 3	**31.24**	**36.36**	**45.45**	**54.54**	**72.7**
		Zone 4	**55.31**	**65**	**85**	**95**	**95**
	6	20 cameras	58.17	66.66	81.81	87.87	93.93
		Zone 1	63.61	66.66	83.33	83.33	83.33
		Zone 2	49.75	37.5	50	56.25	81.25
		Zone 3	37.95	36.36	72.72	72.72	81.81
		Zone 4	58.49	75	90	95	100
	9	20 cameras	59.23	69.69	87.87	87.87	90.9
		Zone 1	57.89	50	66.66	66.66	83.33
		Zone 2	62.47	50	62.5	68.75	75
		Zone 3	53.54	54.54	54.54	72.72	72.72
		Zone 4	74.08	75	85	95	100
	12	20 cameras	53.9	66.6	81.81	87.87	96.96
		Zone 1	53.21	50	83.33	100	100
		Zone 2	51.9	43.75	62.5	75	87.5
		Zone 3	38.12	54.54	54.54	63.6	72.7
		Zone 4	53.76	75	85	85	90
	15	20 cameras	46.46	69.69	81.81	90.9	90.9
		Zone 1	51.75	66.66	66.66	83.33	83.33
		Zone 2	57.47	56.25	68.75	81.25	87.5
		Zone 3	23.211	27.27	36.36	54.54	72.7
		Zone 4	48.15	65	85	90	95

TABLE 6. mAP and rank-k score for ResNet50+Triplet Loss+Re-Ranking.

Image Dimension	P	Mode	mAP	rank-1	rank-5	rank-10	rank-20
128 X 128	3	20 cameras	65.3	72.7	84.8	84.8	93.9
		Zone 1	70	83	83	83	83
		Zone 2	55.6	37.5	68	87	87
		Zone 3	25	36	36	36	54
		Zone 4	49	55	90	90	90
	6	20 cameras	56.4	51.51	69.69	81.81	93.93
		Zone 1	61.1	50	66.6	66.6	100
		Zone 2	53.68	43.75	50	68.75	87.5
		Zone 3	39.71	36.36	45.45	45.45	63.63
		Zone 4	53.3	40	70	85	90
	9	20 cameras	53.32	54.54	72.72	78.78	87.87
		Zone 1	49.2	50	50	66.6	83.3
		Zone 2	47.9	37.5	43.7	50	68.75
		Zone 3	31.72	27.27	45.45	45.45	63.63
		Zone 4	55.1	50	80	85	95
	12	20 cameras	47.7	51.5	60.6	72.7	84.84
		Zone 1	51.43	50	50	83.3	83.3
		Zone 2	39.55	37.5	50	68.75	75
		Zone 3	15.62	18.18	27.27	45.45	54.54
		Zone 4	40	45	70	75	85
	15	20 cameras	45.97	48.48	66.66	75.75	81.81
		Zone 1	46.6	50	50	50	66
		Zone 2	40.92	37.5	37.5	43.7	62.5
		Zone 3	28.44	18.18	36.36	36.36	54.54
		Zone 4	47.48	45	75	85	85
224 X 224	3	**20 cameras**	**77.07**	**84.84**	**93.9**	**100**	**100**
		Zone 1	**80.6**	**83.3**	**83.3**	**83.3**	**100**
		Zone 2	**74.4**	**62.5**	**75**	**87.5**	**93.7**
		Zone 3	**39**	**45.45**	**45.45**	**45.45**	**63.63**
		Zone 4	**77.89**	**90**	**100**	**100**	**100**
	6	20 cameras	65.93	63.63	81.81	84.84	90.9
		Zone 1	60.9	66.66	66.66	66.66	66.66
		Zone 2	65.1	62.5	68.75	68.75	81.25
		Zone 3	54.19	54.54	54.54	54.54	63.63
		Zone 4	62.29	50	80	85	90
	9	20 cameras	66.36	63.63	78.78	84.84	96.96
		Zone 1	63.71	50	66.6	83.33	100
		Zone 2	66.49	50	68.75	75	87.5
		Zone 3	50.83	45.45	54.54	63.63	81.81
		Zone 4	57.13	55	75	80	95
	12	20 cameras	63.56	57.57	72.72	75.75	96.96
		Zone 1	52.49	50	50	66.66	83.33
		Zone 2	61.89	50	62.5	62.5	81.25
		Zone 3	48.08	54.54	54.54	54.54	63.63
		Zone 4	59.11	45	65	70	95
	15	20 cameras	64.98	75.75	93.93	93.93	93.93
		Zone 1	60.2	66.66	66.66	66.66	83.3
		Zone 2	61.76	62.5	68.7	75	75
		Zone 3	40.816	45.45	45.45	54.54	63.63
		Zone 4	61.9	70	95	95	100

undertaken to perform re-identification, a post set of experiments such as Triplet loss and Triplet Loss with L2 regularization are implemented with and without the re-ranking strategy.

For each of the experiments conducted the computed scores are tabulated in the Tables 4,5,6,7,8 respectively. Also Figure 6 illustrates the top-10 match for different query identities observed across entire 20 surveillance cameras.

V. ABLATION STUDY

An analysis of vehicle re-identification algorithms performed across four zones is examined for different sets of experiments. For each experiment, the parameter P (total vehicle identities considered in a batch) [81] is varied in multiples of 3 to study the re-identification performance.

1) **Minimal Setup:** The re-identification network is trained using categorical cross-entropy loss as a minimum setup. During the inference, the presence of query vehicle images is determined by verifying it with vehicle identity images appearing in each zone. Accordingly, the metrics are calculated. The network trained with input dimension 224 × 224 generates better scores than with the network trained with input dimension of 128 × 128 as shown in Table 4. An mAP of 59.82% is obtained during the inference of query images across entire surveillance cameras. An mAP score of 59%, 41.01%, 13.21%, and 55% are obtained while inferring the presence of query vehicle instances for each of the chosen four zones.
2) **ResNet50+Triplet Loss:** The addition of triplet loss enables the re-identification network to distinguish positive and negative labeled vehicles using the generated feature embeddings. The vehicle re-identification experiment with backbone ResNet50 trained with triplet loss is shown in Table 5. It is observed that re-identification performance is significantly better than the network trained with minimal setup. The performance of the vehicle re-identification for a network trained with triplet loss gives better results for inputs of dimension 224 × 224 for a chosen parameter $P = 3$.

TABLE 7. mAP and rank-k score for ResNet50+Triplet Loss+L2 regularization.

Image Dimension	P	Mode	mAP	rank-1	rank-5	rank-10	rank-20
128 × 128	3	20 cameras	48.16	60.6	87.87	93.93	93.93
		Zone 1	55.44	66.6	83.3	83.3	83.3
		Zone 2	48.13	43.75	56.25	68.75	81.25
		Zone 3	31.3	36.36	45.45	45.45	63.63
		Zone 4	46.17	55	80	90	95
	6	20 cameras	50.36	54.54	81.8	90.9	93.93
		Zone 1	51	50	66.6	66.6	66
		Zone 2	53.3	50	50	56.62	62.5
		Zone 3	35.61	36.36	45.45	45.45	63.63
		Zone 4	48.49	45	65	90	100
	9	20 cameras	47.7	54.54	78.78	87.87	90.9
		Zone 1	51.66	66.66	66.66	83.33	83.33
		Zone 2	52.31	43.75	68.7	81.25	81.25
		Zone 3	34.99	27.27	63.63	72.72	72.72
		Zone 4	46.16	60	80	85	90
	12	20 cameras	41.1	48.48	75.75	87.8	90.9
		Zone 1	44.25	50	66.66	66.66	83.33
		Zone 2	57.15	62.5	62.5	81.25	81.25
		Zone 3	34.811	54.54	63.63	63.63	63.63
		Zone 4	45.5	55	80	85	95
	15	20 cameras	47.2	51.51	72.7	78.78	84.84
		Zone 1	52.52	50	66.66	66.66	83.33
		Zone 2	43.9	31.25	31.25	50	68.7
		Zone 3	35.61	36.4	45.45	45.45	45.45
		Zone 4	51.9	45	75	85	90
224 × 224	3	20 cameras	36.82	36.36	54.54	63.63	72.72
		Zone 1	60.9	83.33	83.33	100	100
		Zone 2	39.38	31.25	43.75	68.75	81.25
		Zone 3	15.92	18.18	36.36	36.36	45.45
		Zone 4	46.7	65	90	95	95
	6	**20 cameras**	**61.47**	**69.69**	**87.87**	**93.93**	**96.96**
		Zone 1	**66.51**	**66.66**	**66.66**	**83.33**	**100**
		Zone 2	**58.73**	**50**	**62.5**	**75**	**81.25**
		Zone 3	**32.75**	**36.36**	**54.54**	**63.63**	**72.72**
		Zone 4	**58.83**	**75**	**90**	**95**	**100**
	9	20 cameras	56.39	69.69	75.75	81.81	90.9
		Zone 1	56.2	66.66	83.33	100	100
		Zone 2	59.35	50	68.75	81.25	93.75
		Zone 3	41.47	45.45	54.54	54.54	54.54
		Zone 4	56.5	60	90	90	100
	12	20 cameras	55.82	66.66	78.78	84.84	90.9
		Zone 1	56.73	66.66	66.66	83.33	83.33
		Zone 2	56.5	50	56.2	75	75
		Zone 3	29.72	27.27	54.54	54.54	63.63
		Zone 4	51.94	70	80	85	95
	15	20 cameras	56.86	63.63	81.81	90.9	93.9
		Zone 1	57.65	63.63	84.84	87.87	93.9
		Zone 2	49.29	43.7	62.5	62.5	81.25
		Zone 3	38.46	45.45	54.54	54.54	63.63
		Zone 4	56.76	60	85	95	100

Concerning the above parameters, an mAP of 62.37% is obtained when the query image is looked at across the 20 surveillance cameras. Similarly, for each of the four zones, an mAP of 71.27%, 65.97%, 31.24%, and 55.31% is obtained.

3) **ResNet50+TripletLoss+Re-ranking:** Re-ranking is widely used to increase the re-identification performance to obtain a better rank-k score for a given query image. It is used to improve the generated ranking results for a given query. Using the re-ranking technique for ResNet50 trained with the triplet loss improves re-identification results by a significant percent. For the same parameters observed in the previous experiment, i.e., ResNet50+Triplet Loss, an mAP of 77.07% is obtained considering the entire surveillance cameras for re-identification (Table 6). Also, an mAP of 80.6%, 74.4%, 39%, and 77.89% is obtained when the query images are inferred to each of the four zones.

TABLE 8. mAP and rank-k score for ResNet50+TripletLoss+L2 Regularization+Re-ranking.

Image Dimension	P	Mode	mAP	rank-1	rank-5	rank-10	rank-20
128 × 128	3	20 cameras	61.17	60.6	75.75	90.9	96.96
		Zone 1	64.66	50	83.3	100	100
		Zone 2	43.75	62.5	75	81.25	57.1
		Zone 3	32.12	36.36	36.36	45.45	54.54
		Zone 4	56.63	50	70	85	90
	6	20 cameras	56.64	54.54	75.75	78.78	96.96
		Zone 1	47.33	50	50	66.6	100
		Zone 2	49.6	37.5	43.7	56.25	75
		Zone 3	40.36	36.36	45.45	45.45	63.63
		Zone 4	55.2	50	85	85	95
	9	20 cameras	57.68	57.57	81.81	87.87	90.9
		Zone 1	55.82	50	50	50	83.3
		Zone 2	56.37	37.5	56.25	68.75	87.5
		Zone 3	32.08	27.27	45.45	45.45	63.63
		Zone 4	52.38	55	80	90	90
	12	20 cameras	52.55	54.54	69.69	78.78	84.84
		Zone 1	50.04	50	50	50	83.33
		Zone 2	45.42	37.5	43.75	50	62.5
		Zone 3	41.54	45.45	45.45	54.54	63.63
		Zone 4	51.09	50	75	90	90
	15	20 cameras	50.78	60.6	78.78	81.81	84.84
		Zone 1	43.9	50	50	50	66.6
		Zone 2	52.62	43.7	56.2	62.5	68.7
		Zone 3	26.48	27.27	36.36	36.36	54.54
		Zone 4	48.7	60	80	85	85
224 × 224	3	20 cameras	75.5	75.5	87.87	93.93	96.96
		Zone 1	86.5	83	83	83	100
		Zone 2	75.13	62.5	81.25	87.5	93.7
		Zone 3	42.6	36.36	45.45	54.54	63.63
		Zone 4	70.176	75	85	90	95
	6	20 cameras	66.63	69.69	90.9	93.93	96.96
		Zone 1	72.81	66	100	100	100
		Zone 2	67.19	56.25	81.25	81.25	93.75
		Zone 3	42.7	36.36	54.54	54.54	72.72
		Zone 4	60.67	70	95	100	100
	9	**20 cameras**	**77.22**	**75.75**	**84.84**	**90.9**	**100**
		Zone 1	**82.16**	**100**	**100**	**100**	**100**
		Zone 2	**69.1**	**62.5**	**68.75**	**68.7**	**93.75**
		Zone 3	**66.5**	**63.6**	**72.72**	**72.72**	**72.72**
		Zone 4	**75.76**	**70**	**80**	**90**	**100**
	12	20 cameras	63.7	66.66	78.78	78.78	90.9
		Zone 1	53.42	66.6	83.33	83.33	83.33
		Zone 2	66.12	56.25	68.75	68.75	75
		Zone 3	44.85	36.36	63.63	63.63	72.7
		Zone 4	59.3	55	80	80	95
	15	20 cameras	68.76	69.69	87.87	90.9	100
		Zone 1	66.39	83.33	83.33	83.33	100
		Zone 2	63.8	62.5	75	75	87.5
		Zone 3	53.7	54.54	63.63	63.63	72.7
		Zone 4	64.56	55	85	90	100

TABLE 9. Overall summary of top vehicle re-identification results for different experiments.

Image Dimension	Experiments	P	Mode	mAP	rank-1	rank-5	rank-10	rank-20
224 X 224	Minimal setup	–	20 Cameras	59.82	66	83	83	83
			Zone 1	59	66	83	83	100
			Zone 2	41.01	37.5	56.25	87.5	93.75
			Zone 3	13.21	18.18	27.27	27.27	36.36
			Zone 4	55	90	90	90	90
	ResNet50+ Triplet Loss	3	20 Cameras	62.37	69.69	84.84	93.9	93.9
			Zone 1	71.27	66.6	100	100	100
			Zone 2	65.97	62.5	81.25	93.7	93.7
			Zone 3	31.24	36.36	45.45	54.54	72.7
			Zone 4	55.31	65	85	95	95
	ResNet50+ Triplet Loss+ Re-ranking	3	20 Cameras	77.07	84.84	93.9	100	100
			Zone 1	80.6	83.3	83.3	83.3	100
			Zone 2	74.4	62.5	75	87.5	93.7
			Zone 3	39	45.45	45.45	45.45	63.63
			Zone 4	77.89	90	100	100	100
	ResNet50+ Triplet Loss+ L2 Regularization	6	20 Cameras	61.47	69.69	87.87	93.93	96.96
			Zone 1	66.51	66.66	66.66	83.33	100
			Zone 2	58.73	50	62.5	75	81.25
			Zone 3	32.75	36.36	54.54	63.63	72.72
			Zone 4	58.83	75	90	95	100
	ResNet50+ Triplet Loss+ L2 Regularization+ Re-ranking	9	**20 Cameras**	**77.22**	**75.75**	**84.84**	**90.9**	**100**
			Zone 1	**82.16**	**100**	**100**	**100**	**100**
			Zone 2	**69.1**	**62.5**	**68.75**	**68.7**	**93.75**
			Zone 3	**66.5**	**63.6**	**72.72**	**72.72**	**72.72**
			Zone 4	**75.76**	**70**	**80**	**90**	**100**

4) **ResNet50+TripletLoss+L2 Regularization:** Experiments with L2 regularization are also conducted for vehicle re-identification. The regularizer penalizes

the network to avoid overfitting and allow a better generalization of vehicles while learning the vehicle representation. The re-identification scores, when compared for both image dimension, it is observed from Table 7, for input dimension of 224 × 224 with triplet loss parameter $P = 6$, an mAP of 61.47% is obtained when considering complete surveillance camera information to infer the presence of query images. Similarly, an mAP of 66.51%, 58.73%, 32.75%, and 58.83% is obtained for each of the four zones.

5) **ResNet50+TripletLoss+L2 Regularization+Re-ranking:**
Similar to the experiment (ResNet50+TripletLoss+re-ranking), the use of re-ranking notably improves the mAP scores. From the Table 8 for the network trained with an image dimension of 224 × 224 with the batch hard triplet loss parameter, $P = 9$ yields a mAP of 77.22% for inferring the query images across 20 surveillance cameras. Also for each zone, the mAP of 82.16%, 69.1%, 66.5%, and 75.76% are obtained.

TABLE 10. Comparing average execution time for re-identification of vehicles across zones for ResNet50+TripletLoss+L2 Regularization+Re-ranking.

Mode	Number of Keyframes	Average re-identification time for a query (seconds)
20 Cameras	983	0.862
Zone 1	138	0.297
Zone 2	341	0.428
Zone 3	129	0.211
Zone 4	375	0.566

The re-identification scores in the above experiments were significantly better for the network trained with an image dimension of 224 × 224 . The network trained with smaller image dimensions, i.e., 128 × 128 , cannot learn the precise vehicular representations required to distinguish the vehicles during inference. Also, while propagating through ResNet50 architecture to train an image of 128×128, the feature embedding generated at deeper layers is of lower spatial dimension than the network trained with an image of 224 × 224. It leads to a poor generalization of vehicle representation required to rank the gallery vehicle information for a given query image during inference.

It can be observed that the performance of vehicle re-identification is significantly less for zone 3 compared to other zones. From inferring the surveillance camera data particular to zone 3, it is observed that zone 3 has fewer vehicle movements than the other three zones. Due to limited vehicle movements, a vehicle that appeared in zone 3 traveled at a higher speed resulting in limited vehicle data acquisition by surveillance cameras in zone 3. Hence the pre-processing step of identifying keyframes results in the generation of fewer frames (Figure 7) with the appearance of identical vehicles required to perform re-identification. Hence, determining a given probe image in zone 3 that contains limited vehicle instances appearing in keyframes results in lower re-identification scores. Compared to other zones, vehicles appearing in zone 3 are observed from different perspectives making the network difficult to distinguish similar vehicle feature embeddings observed in zone 3. It is illustrated in Figure 8. Due to these variations in viewpoints that are significantly high in zone 3 compared to other zones, it is observed that the vehicle re-identification algorithm is dependent on the location/placement of surveillance cameras.

The minimal setup yields a lower performance among the different experiments conducted to perform re-identification across zones. Here the re-identification network is trained using categorical cross-entropy loss that performs a pairwise match. However, the network fails to obtain a match for a given identical query if other cameras observe the query vehicle image from different viewpoints. For those experiments that are trained using triplet loss, the network's performance is influenced by the parameter P. The performance of the re-identification network gave lower results when the parameter P was large. For a larger value of P, in each iteration, triplet loss considers P vehicle identities to train the triplet loss network. Since the number of vehicle identities considered for the experimentation is small in number, the triplets generated for each batch to train the network may contain overlapping identical vehicle instances. The use of L2 regularization for re-identification (Table 7) resulted in satisfactory scores when compared with the network trained solely on triplet loss (Table 5). It can be observed that for larger values of P (i.e. $P = 12$ and $P = 15$ for image dimension 224 × 224 in Table 5 and Table 7) network trained with L2 regularization yields a better re-identification score. The regularization factor penalizes the network from overlearning the same representation of vehicles when a batch of images is fed to the network for a higher value of P. Experiments with the re-ranking approach as a post-optimization step yield a higher re-identification score. The network trained with triplet loss with the application of L2 regularization and re-ranking yields a good re-identification score for larger values of P (i.e. $P = 12$ and $P = 15$ in Table 6 and Table 8) than compared with the network trained triplet loss and re-ranking. Table 9 summarizes the overall top vehicle re-identification scores yielded in each of the experiments conducted for vehicle re-identification. As highlighted, vehicle re-identification scores for each experiment were significantly better for images of dimension 224 × 224. For each of the experiments conducted, the performance of vehicle re-identification is found to be superior with ResNet50 backbone network trained with Triplet Loss+L2 regularization and re-ranking (Table 9).

Sometimes, it might be difficult to process the vast amount of surveillance information necessary for re-identifying a vehicle in the real world due to many vehicles moving at the same time. Moreover, looking at all the surveillance videos to re-identify a vehicle of interest can take a great deal of time, thus causing a delay in the re-identification process. Zone specific re-identification can minimize this setback by

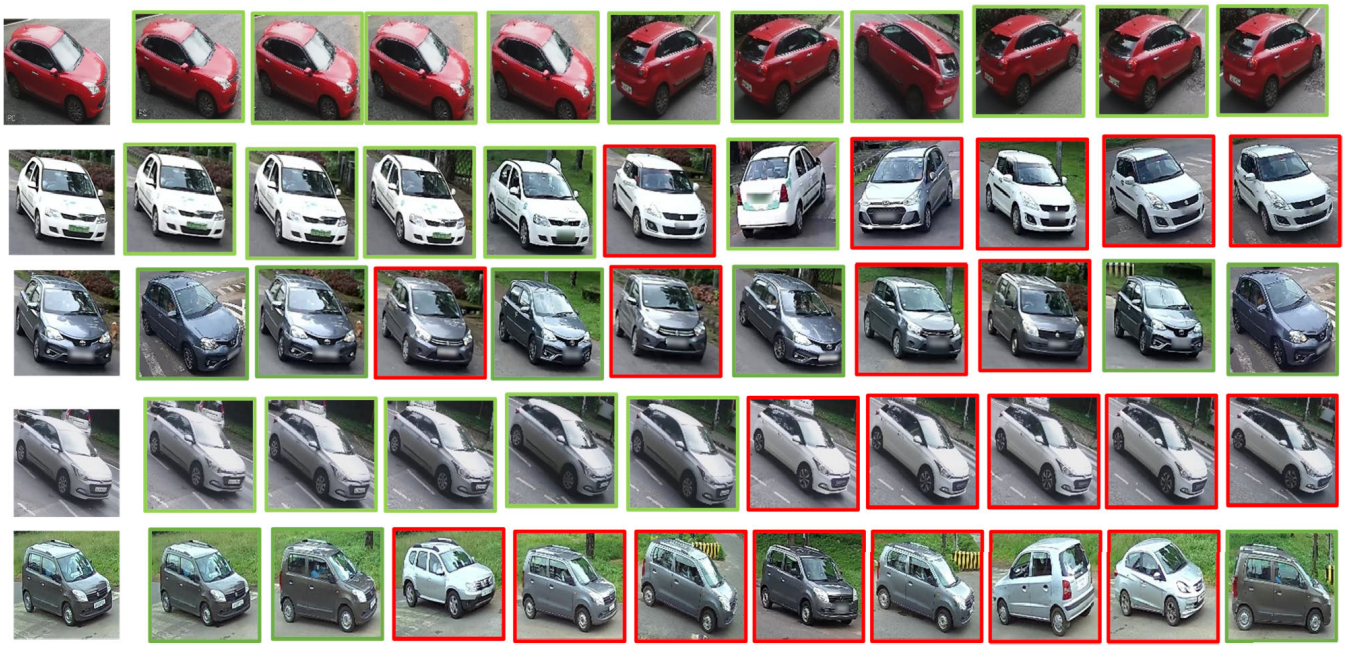

FIGURE 6. Top-10 match for query vehicle identities. Here the first column of every row is a query vehicle image for which the top-10 matches obtained in the gallery set are shown in subsequent columns. Images with green border and red border are the possitve matches and negative matches, respectively.

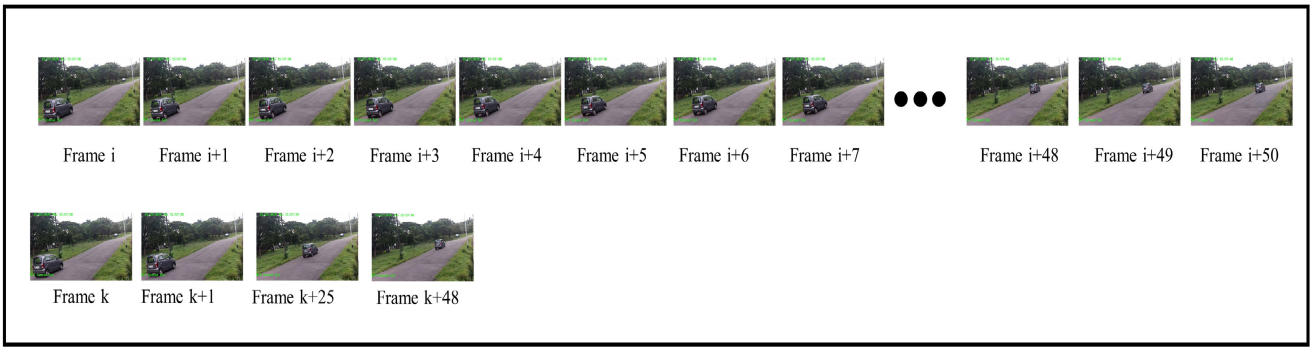

FIGURE 7. Selected keyframes for a given vehicle identity observed in zone 3 surveillance videos. The first row illustrates the raw frames with the presence of the vehicle. The second row highlights only the selected keyframes using shot-boundary detection.

FIGURE 8. Illustration of a vehicle (vehicle with red bounding box) observed at surveillance cameras of zone 3 with significant viewpoint changes.

considering a subset of surveillance data. In a small experiment (Table 10), the average re-identification time for 35 query vehicles is computed when they are re-identified using the entire surveillance cameras (20 cameras). This is determined by estimating the average time required for a given query vehicle that needs to be re-identified across all zones. Also, the average execution time for re-identifying the vehicles in each zone is also shown in Table 10. The model was executed on Intel Xeon Silver 4110 processor clocked at 2.10 GHz, 32 GB of RAM, and a Nvidia GeForce GTX 1080Ti GPU. The average re-identification time is calculated for query vehicles by calculating the time spent extracting features of each query vehicle and features of vehicles appearing in individual zones, and then ranking the gallery vehicles according to their near similarity to the query. Using all surveillance cameras, it takes an average

of 0.862 seconds to re-identify a query vehicle. It can also be observed that among the four zones, zone 3 contains a relatively small number of keyframes. Consequently, the time taken to re-identify the vehicles across zone 3 gallery images is minimal (0.211 seconds) compared to the rest. While considering zone specific surveillance information there is a 65.7% (Zone 1), 50.7% (Zone 2), 75.6% (Zone 3) and 34.6% (Zone 4) of reduction in processing the surveillance information than considering entire surveillance cameras. The significant reduction in the execution time for vehicle re-identification using zones based approach will result in faster identification of vehicles. Indeed, the execution time depends on the total number of frames available as gallery images. The proposed work utilizes shot boundary detection method to identify key-frames thereby reducing the total number of gallery images. This process of eliminating the redundant frames in gallery images reduces the execution time of the vehicle re-identification method. As a result, zone specific surveillance would lead to faster execution, thus cutting down on the time needed to re-identify the vehicles. Using zone specific surveillance, re-identification is performed using a similar methodology as traditional re-identification; however, only a subset of cameras are processed at a time, allowing for faster vehicle re-identification. Therefore, zone specific re-identification is preferred in ITS over the traditional method of re-identification when designing robust traffic management systems for major metropolitan areas.

VI. CONCLUSION

Intelligent Transportation Systems play a prominent role in providing robust traffic management and security to improve transportation network safety and sustainability. In the event of a traffic breach or the cause of an unforeseen event occurring around an area monitored by a group of surveillance cameras, a strategic decision has to be taken by ITS to avoid any damage to the environment or commuters. This paper has undertaken a study to perform vehicle re-identification algorithms across zones with non-uniform vehicle movements. This study assesses the need for vehicle re-identification with a subset of surveillance cameras in a region exposed to traffic breaches. A dataset has been developed to perform zone-specific vehicle re-identification that comprises 81 vehicle identities observed across 20 surveillance cameras. The surveillance cameras are grouped into four zones to monitor non-uniform traffic movements. The re-identification is conducted using a standard CNN backbone architecture for two different image dimensions using triplet loss, L2 regularization, and re-ranking. The network trained with triplet loss along with L2 regularization and re-ranking was robust in re-identifying vehicles across zones. Re-identifying vehicles incorrectly may occur when there is a significant change in viewpoint and poor placement of certain cameras in a zone where human intervention may be possible. In addition, zone-specific re-identification reduces the execution time for re-identifying vehicles in surveillance systems. Zone specific re-identification is therefore an option for ITS that minimises the major trade-offs that need to be made in comparison to the traditional approach of re-identifying the entire surveillance system.

REFERENCES

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

[2] V. Singh and A. K. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Inf. Process. Agricult.*, vol. 4, pp. 41–49, Mar. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214317316300154

[3] N. Liu, Q. Zhao, N. Zhang, X. Cheng, and J. Zhu, "Pose-guided complementary features learning for Amur tiger re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 286–293.

[4] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[5] L.-Y. Hao, J. Li, and G. Guo, "A multi-target corner pooling-based neural network for vehicle detection," *Neural Comput. Appl.*, vol. 32, no. 18, pp. 14497–14506, Sep. 2020.

[6] A. Holla, U. Verma, and R. M. Pai, "Efficient vehicle counting by eliminating identical vehicles in UAV aerial videos," in *Proc. IEEE Int. Conf. Distrib. Comput., VLSI, Electr. Circuits Robot. (DISCOVER)*, Oct. 2020, pp. 246–251.

[7] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[8] D.-Y. Ge, X.-F. Yao, W.-J. Xiang, and Y.-P. Chen, "Vehicle detection and tracking based on video image processing in intelligent transportation system," *Neural Comput. Appl.*, vol. 35, no. 3, pp. 1–13, 2022.

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

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

[11] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 869–884.

[12] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3235–3243.

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

[14] M. M. M. Pai, V. Mehrotra, U. Verma, and R. M. Pai, "Improved semantic segmentation of water bodies and land in SAR images using generative adversarial networks," *Int. J. Semantic Comput.*, vol. 14, no. 1, pp. 55–69, Mar. 2020.

[15] M. M. M. Pai, V. Mehrotra, S. Aiyar, U. Verma, and R. M. Pai, "Automatic segmentation of river and land in SAR images: A deep learning approach," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, Jun. 2019, pp. 15–20.

[16] S. Shorewala, A. Ashfaque, R. Sidharth, and U. Verma, "Weed density and distribution estimation for precision agriculture using semi-supervised learning," *IEEE Access*, vol. 9, pp. 27971–27986, 2021.

[17] U. Verma, F. Rossant, I. Bloch, J. Orensanz, and D. Boisgontier, "Segmentation of tomatoes in open field images with shape and temporal constraints," in *Pattern Recognition Applications and Methods*. Cham, Switzerland: Springer, 2015, pp. 162–178.

[18] S. Girisha, U. Verma, M. M. M. Pai, and R. M. Pai, "UVid-Net: Enhanced semantic segmentation of UAV aerial videos by embedding temporal information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4115–4127, 2021.

[19] S. Girisha, U. Verma, and R. M. Pai, "Semantic segmentation of UAV aerial videos using convolutional neural networks," in *Proc. IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, Jun. 2019, pp. 21–27.

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

[21] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[22] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2017.

[23] T. Surasak, I. Takahiro, C.-H. Cheng, C.-E. Wang, and P.-Y. Sheng, "Histogram of oriented gradients for human detection in video," in *Proc. 5th Int. Conf. Bus. Ind. Res. (ICBIR)*, May 2018, pp. 172–176.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[25] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Proc. Int. Conf. Image Process.*, vol. 1, 2002, p. 1.

[26] D. Zhang, "Vehicle target detection methods based on color fusion deformable part model," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, pp. 1–6, Dec. 2018.

[27] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of online learning and an application to boosting," in *Proc. 2nd Eur. Conf. Comput. Learn. Theory*. Barcelona, Spain: Springer, Mar. 1995, pp. 23–37.

[28] X. Zhang, N. Zheng, Y. He, and F. Wang, "Vehicle detection using an extended hidden random field model," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1555–1559.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[30] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.

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

[33] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9523–9532.

[34] Y. He, J. Han, W. Yu, X. Hong, X. Wei, and Y. Gong, "City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 576–577.

[35] K.-S. Yang, Y.-K. Chen, T.-S. Chen, C.-T. Liu, and S.-Y. Chien, "Tracklet-refined multi-camera tracking based on balanced cross-domain re-identification for vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3983–3992.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[37] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[39] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[40] B. Benjdira, T. Khursheed, A. Koubaa, A. Ammar, and K. Ouni, "Car detection using unmanned aerial vehicles: Comparison between faster R-CNN and YOLOv3," in *Proc. 1st Int. Conf. Unmanned Vehicle Syst. (UVS)*, Feb. 2019, pp. 1–6.

[41] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.

[42] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[43] M. Wu, Y. Qian, C. Wang, and M. Yang, "A multi-camera vehicle tracking system based on city-scale vehicle re-ID and spatial–temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4077–4086.

[44] Z. Zakria, J. Deng, R. Kumar, M. S. Khokhar, J. Cai, and J. Kumar, "Multiscale and direction target detecting in remote sensing images via modified YOLO-v4," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1039–1048, 2022.

[45] G. Jocher, A. Stoken, and J. Borovec, "ultralytics/yolov5: v4.0—nn.SiLU() activations weights & biases logging PyTorch hub integration," Zenodo, Jan. 2021. [Online]. Available: https://zenodo.org/record/4418161, doi: 10.5281/zenodo.4418161.

[46] Y. Liu, X. Zhang, B. Zhang, X. Zhang, S. Wang, and J. Xu, "Multi-camera vehicle tracking based on occlusion-aware and inter-vehicle information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3257–3264.

[47] N. M. Chung, H. D.-A. Le, V. A. Nguyen, Q. Q.-V. Nguyen, T. D.-M. Nguyen, T.-T. Thai, and S. V.-U. Ha, "Multi-camera multi-vehicle tracking with domain generalization and contextual constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3327–3337.

[48] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[49] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[50] D. N.-N. Tran, L. H. Pham, H.-H. Nguyen, and J. W. Jeon, "City-scale multi-camera vehicle tracking of vehicles based on YOLOv7," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2022, pp. 1–4.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE ICCV*, Oct. 2021, pp. 10012–10022.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 213–229.

[54] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[55] Z. Tang, G. Wang, T. Liu, Y.-G. Lee, A. Jahn, X. Liu, X. He, and J.-N. Hwang, "Multiple-kernel based vehicle tracking using 3D deformable model and camera self-calibration," 2017, *arXiv:1708.06831*.

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

[57] A. Ayala-Acevedo, A. Devgun, S. Zahir, and S. Askary, "Vehicle re-identification: Pushing the limits of re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA. Piscataway, NJ, USA: IEEE, Jun. 2019, pp. 291–296.

[58] C.-W. Wu, C.-T. Liu, C.-E. Chiang, W.-C. Tu, and S.-Y. Chien, "Vehicle re-identification with the space-time prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 121–128.

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

[60] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3973–3981.

[61] J. Sochor, J. Špaňhel, and A. Herout, "BoxCars: Improving fine-grained recognition of vehicles using 3-D bounding boxes in traffic surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 97–108, Jan. 2019.

[62] A. Zheng, X. Lin, J. Dong, W. Wang, J. Tang, and B. Luo, "Multi-scale attention vehicle re-identification," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17489–17503, Dec. 2020.

[63] J. Zhang, J. Chen, J. Cao, R. Liu, L. Bian, and S. Chen, "Dual attention granularity network for vehicle re-identification," *Neural Comput. Appl.*, vol. 34, no. 4, pp. 2953–2964, Feb. 2022.

[64] A. Holla, U. Verma, and R. M. Pai, "Enhanced vehicle re-identification for ITS: A feature fusion approach using deep learning," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2022, pp. 1–6.
[65] G. Basar, M. Cetin, and A. P. Nichols, "Comparison of vehicle re-identification models for trucks based on axle spacing measurements," *J. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 517–529, Nov. 2018.
[66] J. Deng, J. Cai, M. U. Aftab, M. S. Khokhar, and R. Kumar, "Visual features with spatio-temporal-based fusion model for cross-dataset vehicle re-identification," *Electronics*, vol. 9, no. 7, p. 1083, Jul. 2020.
[67] K.-T. Nguyen, T.-H. Hoang, T.-N. Le, M.-T. Tran, N.-M. Bui, T.-L. Do, V.-K. Vo-Ho, Q.-A. Luong, M.-K. Tran, T.-A. Nguyen, T.-D. Truong, V.-T. Nguyen, and M. Do, "Vehicle re-identification with learned representation and spatial verification and abnormality detection with multiadaptive vehicle detectors for traffic video analysis," in *Proc. CVPR Workshops*, 2019.
[68] K. Lv, H. Du, Y. Hou, W. Deng, H. Sheng, J. Jiao, and L. Zheng, "Vehicle re-identification with location and time stamps," in *Proc. CVPR Workshops*, 2019, pp. 399–406.
[69] M. Naphade, Z. Tang, M.-C. Chang, D. C. Anastasiu, A. Sharma, R. Chellappa, and S. Wang, "The 2019 AI city challenge," in *Proc. CVPR Workshops*, 2019, pp. 452–460.
[70] M.-C. Chang, C.-K. Chiang, C.-M. Tsai, Y.-K. Chang, H.-L. Chiang, Y.-A. Wang, S.-Y. Chang, Y.-L. Li, M.-S. Tsai, and H.-Y. Tseng, "AI city challenge 2020—Computer vision for smart transportation applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 620–621.
[71] Y. Hou, H. Du, and L. Zheng, "A locality aware city-scale multi-camera vehicle tracking system," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 167–174.
[72] P. Khorramshahi, N. Peri, A. Kumar, A. Shah, and R. Chellappa, "Attention driven vehicle re-identification and unsupervised anomaly detection for traffic understanding," in *Proc. CVPR Workshops*, 2019, pp. 239–246.
[73] S. Teng, S. Zhang, Q. Huang, and N. Sebe, "Viewpoint and scale consistency reinforcement for UAV vehicle re-identification," *Int. J. Comput. Vis.*, vol. 129, no. 2385, pp. 719–735, 2020.
[74] P. Wang, B. Jiao, L. Yang, Y. Yang, S. Zhang, W. Wei, and Y. Zhang, "Vehicle re-identification in aerial imagery: Dataset and approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 460–469.
[75] B. A. Holla, M. M. M. Pai, U. Verma, and R. M. Pai, "Vehicle re-identification in smart city transportation using hybrid surveillance systems," in *Proc. IEEE Region 10 Conf. (TENCON)*, Dec. 2021, pp. 335–340.
[76] X. Ni, L. Fang, and H. Huttunen, "Adaptive L_2 regularization in person re-identification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2020, pp. 9601–9607.
[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[78] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[79] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13001–13008.
[80] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[81] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
[82] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

B. ASHUTOSH HOLLA 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 with the Manipal Institute of Technology. His research interests include object detection, re-identification, and deep learning for computer vision.

M. M. MANOHARA PAI (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering. He is currently a Professor with the Department of Information and Communication Technology, Manipal Academy of Higher Education, Manipal, India. He has rich experience of 31 years as a Professor and has recently received National Technical Teacher's Award (NTTA-2022) from the ministry of education, Govt. of India. He holds seven patents to his credit and has published 131 papers in national and international journals/conference proceedings. He has published two books, guided six Ph.D. and 85 master's theses. 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 the Systems Society of India. He is a principal investigator for multiple industry/government research projects. He has been an Executive Committee Member of the IEEE Bangalore Section and Mangalore Subsection, and the past Chair of the IEEE Mangalore Subsection.

UJJWAL VERMA (Senior Member, IEEE) received the Ph.D. degree in image analysis from Télécom ParisTech, University of Paris-Saclay, Paris, France, and the M.S. (Research) degree in signal and image processing from IMT Atlantique, France. He is currently an Associate Professor and the Head of the Department of Electronics and Communication Engineering, 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 also a Life Member of the Indian Science Congress Association. He was a recipient of the "ISCA Young Scientist Award" (2017–2018) by the Indian Science Congress Association (ISCA), a professional body under the Department of Science and Technology, Government of India. He is also the Co-Lead of 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 also a Guest Editor of Special Stream in IEEE Geoscience and Remote Sensing Letters and a Reviewer for several journals, including IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, and IEEE Geoscience and Remote Sensing Letters. He is also a Sectional Recorder of the ICT Section of the Indian Science Congress Association (2020–2022).

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 total of 30 years of experience in teaching and research. She has published 87 papers in national/international journals/conferences and has guided three Ph.D. 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 the National Doctoral Fellowship from AICTE, Government of India.