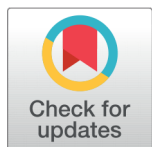


ORIGINAL ARTICLE



OPEN ACCESS

Received: 05/04/2026

Accepted: 13/07/2026

Published: 29/07/2026

Citation: Ilamathi M, Ramakrishnan S (2026) Precise Localization Strategy for Reliable Multi-Vehicle Tracking in Unregulated Road Traffic Based on Enhanced Detection Transformer. Indian Journal of Science and Technology 19(27): 1955-1963. <https://doi.org/10.17485/IJST/v19i27.534>

* **Corresponding author.**

ilamathm@srmist.edu.in

Funding: None

Competing Interests: None

Copyright: © 2026 Ilamathi & Ramakrishnan. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indst.org/))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Precise Localization Strategy for Reliable Multi-Vehicle Tracking in Unregulated Road Traffic Based on Enhanced Detection Transformer

M Ilamathi^{1*}, Sabitha Ramakrishnan²

¹ Department of Computer Science and Engineering with Specialization in Cyber Security, SRM Institute of Science and Technology, Ramapuram, Chennai, India

² Department of Instrumentation Engineering, Madras Institute of Technology, Anna University, India

Abstract

Objectives: To develop a reliable multi-vehicle detection-based tracking framework under unregulated traffic scenario using enhanced vision transformer model that can precisely localize and track the vehicles. **Method:** The proposed framework involves a three-level enhancements, (1) Adaptive frame slicing technique in the input data loader to capture the discriminative features of near and far away vehicles in the video frame (2) Geometric positional encoding to provide clear spatial cues and fusion of features, allowing the model to better separate and detect multi-scale vehicles (3) Dual inferencing of sliced and full frame with DIOU-NMS post-processing technique to remove redundant detections. The proposed framework is evaluated on our custom developed Active Learning based vehicle dataset, "AU-INV-P-PALS", and two public traffic datasets- "ITM-HeTra and Udacity-Fixed small". **Findings:** In comparison with the existing baseline DETR model, the proposed model with SGD optimization showed 9.2% increase in the average precision for IoU = 0.5:0.95 indicating its precise localization ability. The model localizes vehicles that occupy as little as 0.2% of the surveillance video frame area. The observed tracking accuracy on real-time traffic CCTV videos indicates that the model can adapt to unregulated traffic environment and achieved appreciable score of about 85.4%, maintaining the tracking stability. The average F1-score of the enhanced DETR obtained for the test frames indicate the generalization ability of the model across different frames. **Novelty:** The proposed work introduces adaptive frame slicing scheme in the input data loader and geometric positional encoding which enables the detection of faraway vehicles with high accuracy in wide area surveillance imagery. Further, the enhanced model is trained using active learning based custom developed vehicle dataset that significantly reduces the reliance on massive datasets.

Keywords: Deep Learning; Active Learning; Unregulated Road Traffic Control; Vision Transformer

1 Introduction

The advancement of Automatic Traffic Management Systems (ATMS) in developing countries is a critical initiative in addressing the rapid urbanization and enhancing the vision of smart cities. Multi-Vehicle Detection Based Tracking (MVDDBT) is an important component of ATMS and serves as the basis for traffic regulation. Traffic is fairly regulated in most of the developed countries, while in India we can identify some distinct differences such as traffic density, pedestrian movements, animal crossing, unique vehicle types, varying surveillance camera viewpoints and multiple road intersections. Figure 1 shows a wide area surveillance imagery⁽¹⁾ of unregulated road traffic with dense and variable-scale vehicles. The size of the vehicles varies with the distance from the camera, leading to poor localization accuracy when they are far away.

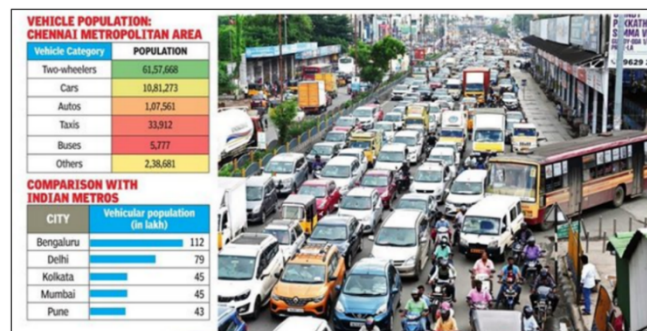


Fig 1. Unregulated traffic of Chennai metropolitan area with dense varying scale vehicle population

However, ATMS technologies that are implemented in developed countries cannot be directly deployed for Indian traffic monitoring and control. Suitable modifications along with adoption of new learning strategies are required for understanding and handling the data-level (aleatoric) and algorithmic (epistemic) uncertainties to ensure effective implementation of ATMS for MVD&T in our country where unregulated traffic environment is prevalent in most of the roads.

The use of deep learning models has received wide interest in DBT and simultaneously perform detection, motion prediction followed by data association which allows them to overcome challenges like occlusions and fast-moving vehicles⁽²⁾. With significant progress in deep learning and transformer models, the motion prediction and data association stages have been merged for tracking purpose. Prominent tracking algorithms can be integrated and the rich visual features extracted by the CNN module are fed into the data association stage which aids the representation of discriminative features, thereby achieving high tracking accuracy^{(3), (4)}.

Convolutional Neural Networks (CNN's) such as Faster RCNN⁽⁵⁾, YOLO series⁽⁶⁾, SSD⁽⁷⁾, RetinaNet⁽⁸⁾, and EfficientDet⁽⁹⁾ have been commonly utilized for vision-based detection task. While CNNs excel in extracting local and hierarchical features, they struggle with complex traffic scenes involving interactions between vehicles or when vehicles are partially hidden owing to occlusion affecting the real-time performance. Recently, Vision Transformers with self-attention mechanisms gained popularity for object-detection applications⁽¹⁰⁾. DETR⁽¹¹⁾ is a milestone in these efforts, and excel at capturing long-range dependencies and global contexts, making them better suited for understanding complex image patterns. However, when the traffic conditions and vehicle appearances vary greatly based on time of the day, camera resolution and camera angle, these transformer-based models do not show good performance in MVD as they are designed for general object detection tasks only. They utilize two-dimensional positions as reference points, disregarding the scale of the targets. This results in a suboptimal detection performance for smaller targets. Also, object queries in the decoder struggle to perceive intricate image content and exhibit discrepancies from the ground truth in the detection stage⁽¹²⁾. Hence, effective domain adaptation and enhancement strategies on top of the base architecture are required in order to detect all-scale vehicles with minimal additional architectural complexity.

To circumvent these challenges, this paper proposes three enhancement schemes over the baseline Detection Transformer model, to achieve multi-vehicle detection-based tracking accurately irrespective of their size within the video frame. The proposed model is trained and validated by utilizing three experimental unregulated traffic environments-based datasets namely, AU-INV-P-PALS generated using a pool-based active learning strategy⁽¹³⁾, IITM-HeTra⁽¹⁴⁾ and Udacity-Fixed small⁽¹⁵⁾.

The major contributions in the proposed work are as follows:

1. Pre-processing the input data using adaptive frame slicing technique to enable the backbone network of transformers to extract the discriminative features of far-way vehicles in the frame.

2. Geometric positional encoding provides clear spatial cues and fusion of features, allowing the model to better separate and detect multi-scale vehicles.
3. Optimal hyperparameter tuning with different optimizers (SGD, AdamW, and RMSProp) along with transfer learning-based model training to enable faster convergence with high detection accuracy.
4. Dual inferencing of sliced and full frame with DIOU-NMS post-processing technique to remove redundant detections
5. Integration of the localization results achieved by enhanced DETR with the efficient tracker algorithms to perform continuous detection-based tracking.
6. The proposed model has been successfully trained and validated on three datasets namely, AU-INV-P-PALS, Udacity-Fixed small and IITM-HeTra datasets. The reliability of the proposed model is tested using real-time CCTV traffic footage.

2 Methodology

The proposed methodology involves a sequence of stages starting from the detection stage, followed by motion prediction to predict the future coordinates of the detected vehicle, and data association between the detected and predicted vehicle positions to maintain track continuity as shown in Figure 2.

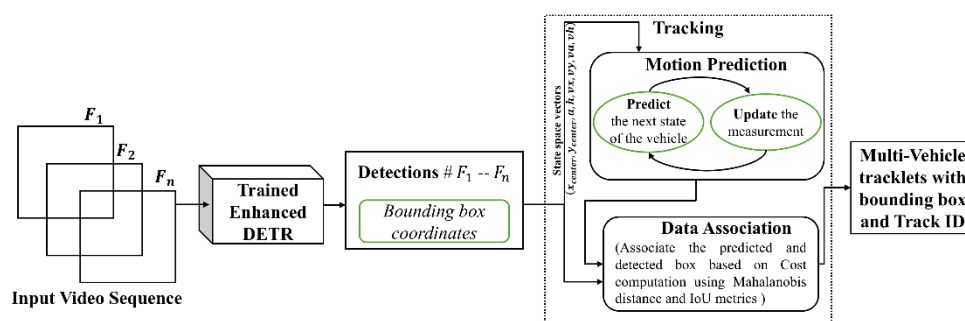


Fig 2. Outline of the proposed detection based multi-vehicle tracking approach

2.1 Proposed enhanced DETR for precise vehicle localization

A schematic of the proposed enhanced Detection Transformer (DETR) for multi-vehicle localization and classification is shown in Figure 3.

2.1.1. Adaptive frame slicing and enhanced positional encoding technique

The original high-resolution video frames are divided into smaller manageable slices. The number of slices is determined based on the three parameters namely, width (W) and height (H) of the original image, side lengths of each slice ($S = w_s \times h_s$) and overlap ratio r , where $0 < r < 1$. A uniform overlap ratio is maintained for all the slices using the iterative search algorithm by creating a grid with intervals of 0.05 in the range $[0, 0.5]$ and adjusting the overlap ratio accordingly.

The pseudocode for the step-by-step computation of the number of slices for a given video frame of 1920x1080 dimension is illustrated below.

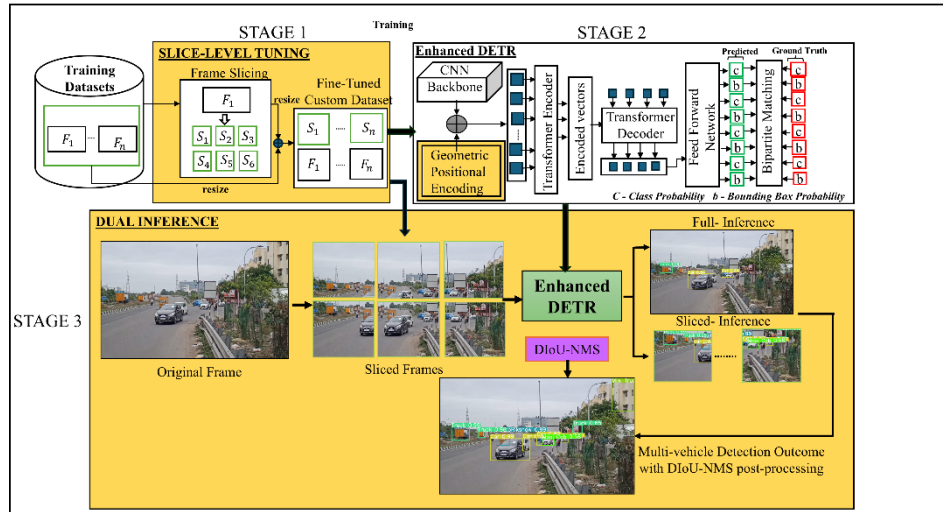


Fig 3. Schematic of the proposed framework multi-vehicle detection using Enhanced DETR

Pseudo Code for adaptive frame slicing:

Input: High Resolution frame with size 1920×1080

Output: Total number of Slices (N_T)

1. Initialize image dimensions:

$W = 1920$

$H = 1080$

2. Determine the optimal Side lengths of each slice across width and height
(This is based on specific requirements or constraints related to the problem)

Side Lengths (w_s and h_s) = Determine_Optimal_SliceSize ()

3. Set the overlapping ratio: $r = 0.2$

4. Compute effective_side_length across width and height using math equation:

$w_s \times (1 - r)$ and $h_s \times (1 - r)$

5. Compute the number of slices along the width:

$N_W = \lceil \frac{W - (w_s \times r)}{w_s \times (1 - r)} \rceil$ and $N_H = \lceil \frac{H - (h_s \times r)}{h_s \times (1 - r)} \rceil$

6. Calculate total number of slices: $N_T = N_W \times N_H$

End

The sliced frame $S \in R^{H \times W \times 3}$ (H , W and C denote the height, width, and number of colour channels respectively), is reduced to a feature map $f \in R^{h_0 \times w_0 \times C}$ that retains the significant details and a 1×1 convolution is applied to the reduced 2D spatial structure to obtain one-dimensional input embeddings. Positional encoding is added to these input embeddings to understand the relative positions of the different elements of the image and is given to the encoder block.

In the proposed work, the absolute encoding technique utilized in the baseline DETR is replaced with geometric positional encoding to adapt the scalability for a fleet of vehicles with varying extrinsic calibrations. Geometric encodings can model relative offsets with a linear operator using a three-dimensional rotation matrix as the operator, which leads to a nuanced understanding of the spatial relationships within an image.

The transformed position is given by

$$P = R(\theta)P \quad (6)$$

The final geometric positional encoding is given by Equation (7).

$$PE_{(i,j)} = (i \cos \theta - j \sin \theta + \Delta i, i \sin \theta + j \cos \theta + \Delta j) \quad (7)$$

This encoding transforms spatial locations in a way that captures relative positional offsets via a learned rotation and translation, enabling the model to dynamically understand spatial relationships.

The feature map extracted with significant information from the input slices is concatenated with the geometric encodings to form the encoder input embedding vectors: query (q), key (k), and value (v). Using these vectors, the attention weight ($W_{i,j}$) for each pixel is computed. The scaled dot product is used to compute the attention weights, as indicated in Equation (8), where, Softmax activation is applied to normalize the weights.

$$W_{i,j} = \sum_{1,m \in N_k(i,j)} \text{softmax}_{l_m}(q_{ij}^T k_{lm}) v_m \quad (8)$$

where $q_{ij} = p_q X_{ij}$, $K_{lm} = p_k X_{lm}$, $v_{lm} = p_v X_{lm}$ denote the linear pixel points and the changes in surrounding pixel points respectively, and $P_q, P_k, P_v \in R^{d \times H \times W}$ are the parameters that the network needs to learn.

Finally, the decoder takes the object queries in parallel with the encoder output and transforms them into output embeddings, which are subsequently decoded into box coordinates by the Feed-Forward Network (FFN). A linear layer accompanies the FFN to assign the most probable class label for each vehicle. After the model infers a set of N predictions the ground and the predicted vehicles are matched using, the bipartite-matching. This approach enables the model to effectively match the predicted vehicles with the ground truth labels.

2.1.2. DIOU-NMS

Dual inference is performed on both the original frame and sliced frames to facilitate the detection of multi-scales vehicle targets. The incorporation of dual inference requires the suppression of false negatives and overlapping boxes on the same vehicle targets. The traditional IoU based suppression techniques account for the overlapping area only and neglects centroid distance. Hence, in the proposed work, Distance IoU based Non-maximum Suppression (DIOU-NMS)⁽¹⁶⁾ is used, which considers both the overlapping area and distance of the center point of the two bounding boxes, which is given by Equation (10).

$$DIOU = IOU - \frac{\rho^2(c1, c2)}{d^2} \quad (9)$$

where, IOU denotes Intersection of Union, d^2 denotes the diagonal distance of the smallest box overlaying the two bounding boxes, and $\rho^2(c1, c2)$ gives the Euclidean distance between the centroids of the two bounding boxes.

2.2 Multi-vehicle tracking

The tracking process is implemented using state-of-the-art tracking algorithms, DeepSORT and ByteTrack⁽¹⁷⁾. The enhanced DETR model generates precise localization results across consecutive video frames, which are subsequently fed into the motion prediction stage. This stage facilitates temporal consistency in vehicle tracking, followed by a data association step that establishes and maintains robust trajectories of moving vehicles. The integration of motion modelling and association mechanisms ensures accurate and continuous vehicle tracking across dynamic scenes. Kalman filter⁽¹⁸⁾ predicts the next state of the bounding box based on the previous state and motion model. The predicted state estimate is given by Equation (11)

$$\hat{S}_{(t|t-1)} = J_t S_{(t-1|t-1)} \quad (10)$$

where, $\hat{S}_{(t|t-1)}$ is the predicted state at time t and J_t is the state transition matrix computed at each time t to handle the non-linearity.

Once the Kalman Filter predicts the bounding box and updates it with the detection, the next step is to associate the predicted bounding box with the detected bounding boxes from enhanced DETR. Further, the data association stage is carried out in two steps. In the first association step, the high-confidence detection boxes are matched across the detection boxes from previous frames based on the pre-set matching threshold. In the second association step, the matching threshold is reset with a lower value to match the low-confidence detection boxes with the unmatched detected boxes from previous frames. The tracklets (bounding boxes) that are lost are retained for a certain number of frame duration and included back to the active tracklets. This allows the tracker to recover the missed tracklets that have been lost due to severe occlusion that makes the targets disappear for a short number of frames. This special feature of ByteTrack algorithm helps in handling complex traffic patterns in unregulated traffic scenarios.

2.3 Implementation settings and parameter optimization

The detection-based tracking framework is implemented using Python libraries, supervision v0.3.0, transformers v4.35.2, and PyTorch-lightning v2.1.0. Google Colab Pro featuring 32GB RAM, Tesla T4 cloud GPU, is utilized for training and testing.

The enhanced DETR model is trained using transfer learning on AU-INV-P-PALS, IITM-HeTra and Udacity fixed small datasets. AU-INV-P-PALS is a custom developed dataset recorded at 30 frames per second (fps) under an intractable traffic scenario using a stationary camera positioned at different angles (facing the road, from the top, and sideways) and at various times of the day (morning, noon, and evening). The ratio of training to test batches is randomly split to 70:30⁽¹⁹⁾ and trained for 100 epochs, utilizing six encoder and decoder layers. Three optimization techniques, AdamW, SGD, and RMSProp, are chosen for optimization and the optimal technique is identified by comparing their training and validation loss curves pertaining to bounding box loss and classification loss.

3 Results and discussion

3.1 Effect of parameter optimization

The training and validation loss convergences pertaining to the bounding box and classification using the three optimizers on the AU-INV-P-PALS dataset is shown in Figure 4. The SGD optimizer achieved fast convergence with minimal loss for 100 training epochs. After extensive experimentation, enhanced DETR trained on custom-made AU-INV-P-PALS with learning rate = $1e^{-4}$, weight decay = $1e^{-5}$, momentum = 0.9 and backbone learning rate = $1e^{-5}$ during training achieves appreciable average precision of 85.6% for IoU at 0.5 and average recall of 87.1% ensuring the detection of all vehicle regions with dimensions ranging from 60×60 to 1024×1024 which is desirable for real-time implementation.

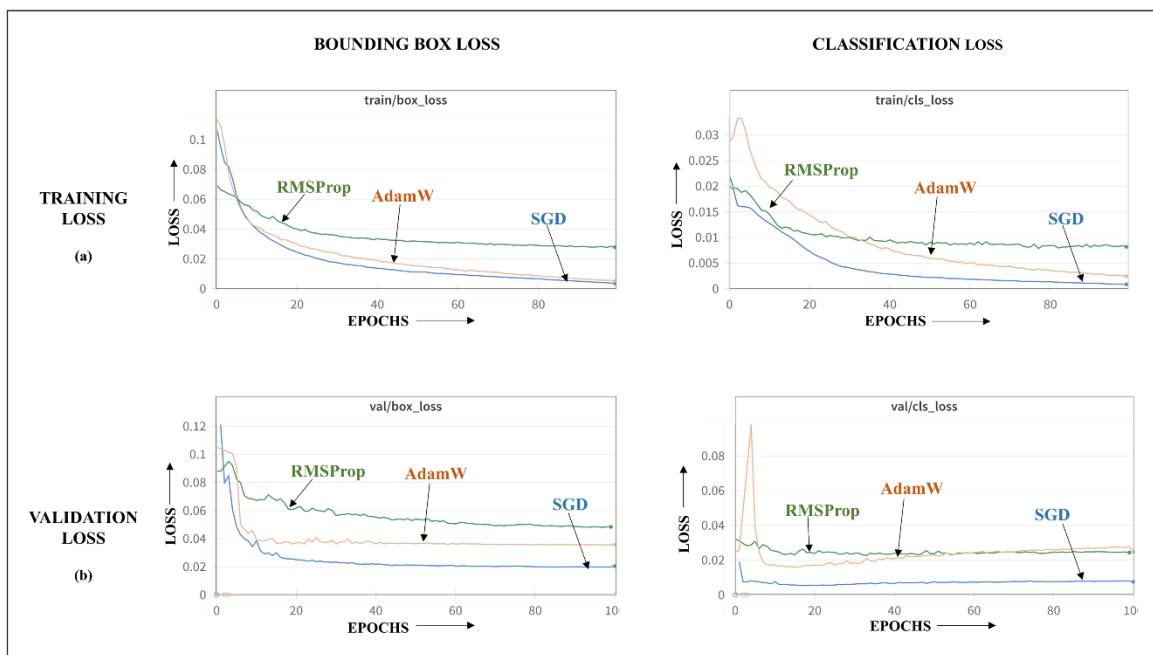


Fig 4. Loss curves of Enhanced DETR on AU-INV-P-PALS dataset (a) Training loss and (b) Validation loss

3.2 Quantitative and qualitative validation of enhanced DETR with existing models

Table 1 presents the comparative analysis of the proposed models with existing vehicle detection models proposed in the recent literature for vehicle detection using comprehensive performance measures namely, mAP@0.5:0.95, AR₁₀₀ (Average Recall per 100 detections) and F1-score. The higher accuracy values computed for IoU = 0.5:0.95 compared with other existing models indicate that the proposed model precisely localizes the vehicles across all scales and aspect ratios present in the unregulated traffic.

Table 1. Comparative analysis of Enhanced DETR with existing vehicle detection models

Model	Datasets	mAP@ 0.5:0.95	AR ₁₀₀	F1-Score	Inference Time (s)
Enhanced DETR (Proposed)	AU-INV-P-PALS	76.1	73.9	74.4	0.0201
DETR	AU-INV-P-PALS	69.7	67.3	68.3	0.0171
Faster R-CNN	AU-INV-P-PALS	66.8	64.8	67.3	0.14
Mask R-CNN	AU-INV-P-PALS	66.3	60.6	60.9	0.2
EfficientDet	AU-INV-P-PALS	62.4	60.9	61.6	0.016
YOLOv5s	AU-INV-P-PALS	63.6	61.6	62.8	0.0143
Enhanced DETR (Proposed)	IITM-HeTra	74.4	72.8	73.0	0.0201
DETR	IITM-HeTra	69.5	67.5	69.9	0.0171
Faster R-CNN	IITM-HeTra	66.1	65.3	67.6	0.14
YOLOv3	IITM-HeTra	66.2	62.7	69.9	0.021
YOLOv4	IITM-HeTra	64.1	73.3	73.4	0.0192
EfficientDet	IITM-HeTra	67.9	67.2	67.0	0.016
Enhanced DETR (Proposed)	Udacity	67.4	70.1	68.7	0.0201
DETR	Udacity	61.7	68.6	68.1	0.0171
YOLOv4	Udacity	58.9	66.1	62.2	0.0192
YOLOv5s	Udacity	59.0	62.9	60.8	0.0143
CMS-YOLO	Udacity	61.7	63.5	64.4	0.0143

Further, to emphasize the balance between precision and recall score, F1 score is computed for all video frames present in the test set of the respective datasets through the scatter plot shown in Figure 5.

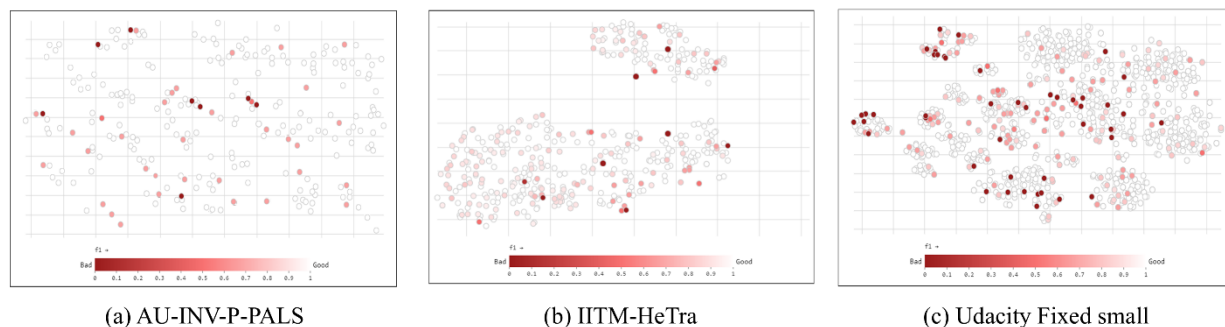


Fig 5. Vector Analysis of F1-score on test sets of (a) AU-INV-P-PALS (b) IITM-HeTra (c) Udacity Fixed small

Each dot represents one frame in the test set and the nodes are colored by their F1-scores, lighter shade indicates better F1 score, and darker shade indicate poorer F1 score. The higher concentration of lighter dots indicates superior F1-scores across most test set frames, which proves that the proposed model not only detects vehicles with high accuracy but also minimizes false positives.

3.3 Multi-vehicle tracking inference

Table 2 presents the comparative tracking results attained by enhanced DETR. Two real-time traffic videos recorded at 30 fps are used as test videos and the performance is analyzed using Multi-Object Tracking Accuracy (MOTA), Identity F1 score (IDF1), and the processing speed (fps).

The enhanced DETR-ByteTrack achieved approximately 85% tracking accuracy while operating at a processing speed of around 50 frames per second (fps). In comparison, the enhanced DETR model alone processes approximately around 49-50 fps, whereas the baseline model achieves a higher processing speed of around 60 fps. However, considering that real-time surveillance videos typically operate within a range of 30 to 50 fps, the enhanced DETR-ByteTrack framework remains well-suited for real-time surveillance applications.

Table 2. Tracking results of the detection-based tracking algorithm

Test Data	DBT Models	MOTA (%)	IDF1 (%)	Processing Speed (fps)
Real-time CCTV Video 1	Baseline DETR-DeepSORT	72.6	74.2	58.8
	Baseline DETR-ByteTrack	73.2	74.1	59.0
	Enhanced DETR-DeepSORT	83.9	84.5	49.2
	Enhanced DETR-ByteTrack	85.3	86.6	49.2
Real-time CCTV Video 2	Baseline DETR-DeepSORT	72.1	73.5	58.1
	Baseline DETR-ByteTrack	72.9	73.4	58.8
	Enhanced DETR-DeepSORT	83.4	84.2	49.0
	Enhanced DETR-ByteTrack	85.1	85.9	49.2

A qualitative analysis of Enhanced DETR is performed using the real-time CCTV footage captured from Chennai City and the multi-vehicle tracking inferences is illustrated in Figure 6.


Fig 6. Multi-vehicle tracking on Video 1 and Video 2 with unique ID using Enhanced DETR-ByteTrack algorithm

It is observed that the enhanced DETR, due to its attention mechanism, helps the tracker to efficiently associate each vehicle across frames, ensuring long tracklet without any misses from the beginning of the video sequence till the vehicle disappears from the frame.

4 Conclusion

This study presents a precise localization based multi-vehicle tracking framework to serve as a promising option for establishing an Advanced Traffic Management System (ATMS) in developing nations. The proposed framework with slicing-aided frame level tuning achieved high-accuracy, specifically excelling at identifying distant objects in wide-area surveillance. Using SGD optimization, the model reaches an accuracy of 85.6% with inference speed of 20.3 milliseconds, making it ideal for real-time processing. The high mAP@0.5:0.95 score compared with existing models reflects high localization and detection accuracy. The comparative increase as compared to the public datasets is due to the reason that the proposed model utilizes active learning based custom dataset which enables the model to detect vehicles well at both lower IoU threshold and at higher IoU thresholds. The active learning based custom dataset that has been used for training the proposed models greatly contributes towards enhanced performance and reduced training time. The tracking efficiency is further demonstrated through real-time validation

on CCTV footage, where it successfully localizes and tracks vehicles occupying as little as 0.2% of the total frame area under complex traffic environments.

5 Acknowledgements

The authors thank Greater Chennai Corporation for their support throughout the real-time traffic data collection process.

Author details: Dr.M Ilamathi: Assistant Professor; Dr. Sabitha Ramakrishnan: Professor.

Contributory statement: Conceptualization of study problem: M. Ilamathi. & Sabitha Ramakrishnan., Experimental Design: M. Ilamathi., Algorithm: M. Ilamathi., Data generation: M. Ilamathi., Original Drafting: M. Ilamathi., Research Guidance, Editing & Reviewing: Sabitha Ramakrishnan.

References

- 1) of India TT. People in suburbs are buying more bikes, cars than in core city area. 2023 Jan. Available from: <https://timesofindia.indiatimes.com/city/chennai/more-cars-and-bikes-drive-out-to-suburbs-in-chennai/articleshow/97158378.cms>.
- 2) Zhao G, Meng F, Yang C, Wei H, Zhang D, Zheng Z. A review of object tracking based on deep learning. *Neurocomputing*. 2025;651:130988. <https://doi.org/10.1016/j.neucom.2025.130988>.
- 3) Liao N, Ran C, Zhu M, Yan H, Peng B, Lu Y. MCF-DETR: Lightweight multi-scale feature fusion transformer for small object detection in UAV imagery. *Image and Vision Computing*. 2026;174:106098. <https://doi.org/10.1016/j.imavis.2026.106098>.
- 4) Zhang Y, Li L, Wu G, Tang B, Zhao W. TransSTC: Transformer tracker meets efficient spatial-temporal cues. *Pattern Recognition*. 2026;172(Part A):112303. <https://doi.org/10.1016/j.patcog.2025.112303>.
- 5) Ahmed SH, Raza M, Mehdi SS, Rehman I, Kazmi M, Qazi SA. Faster RCNN based Vehicle Detection and Counting Framework for Undisciplined Traffic Conditions. 2021 *IEEE 18th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*. 2021. <https://doi.org/10.1109/HONET53078.2021.9615466>.
- 6) Murat AA, Kiran MS. A comprehensive review on YOLO versions for object detection. *Engineering Science and Technology, an International Journal*. 2025;70:102161. <https://doi.org/10.1016/j.jestech.2025.102161>.
- 7) Kumar A, Zhang ZJ, Lyu H. Object detection in real time based on improved single shot multi-box detector algorithm. *EURASIP Journal on Wireless Communications and Networking*. 2020;2020(1):204. <https://doi.org/10.1186/s13638-020-01826-x>.
- 8) Karbouj B, Topalian-Rivas GA, Krüger J. Comparative Performance Evaluation of One-Stage and Two-Stage Object Detectors for Screw Head Detection and Classification in Disassembly Processes. *Procedia CIRP*. 2024;122:527–532. <https://doi.org/10.1016/j.procir.2024.01.077>.
- 9) Tan M, Pang R, Le QV. EfficientDet: Scalable and Efficient Object Detection. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. <https://doi.org/10.1109/CVPR42600.2020.01079>.
- 10) Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;45(1):87–110. <https://doi.org/10.1109/TPAMI.2022.3152247>.
- 11) Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. *Computer Vision – ECCV 2020*. 2020. <https://doi.org/10.48550/arXiv.2005.12872>.
- 12) Liu H, Qin J, Zhang Z, Liang H, Deng H, Xiao S, et al. VeTrack leverages transformer-based appearance modeling and confidence-guided multi-level association for robust multi-vehicle tracking. *Scientific Reports*. 2026;16:2673. <https://doi.org/10.1038/s41598-025-32425-9>.
- 13) Ilamathi M, Ramakrishnan S, Babusankar RK. Proactive hybrid learning framework for real-time multi-vehicle detection in unregulated traffic environments. *Image and Vision Computing*. 2024;147:105081. <https://doi.org/10.1016/j.imavis.2024.105081>.
- 14) Mittal D, Reddy A, Ramadurai G, Mitra K, Ravindran B. Training a deep learning architecture for vehicle detection using limited heterogeneous traffic data. 2018 *10th International Conference on Communication Systems & Networks (COMSNETS)*. 2018. <https://doi.org/10.1109/COMSNETS.2018.8328279>.
- 15) Dataset USDC. Roboflow. 2020 Feb. Available from: <https://public.roboflow.com/object-detection/self-driving-car/3>.
- 16) Wang M, Fu B, Fan J, Wang Y, Zhang L, Xia C. Sweet potato leaf detection in a natural scene based on faster R-CNN with a visual attention mechanism and DIOU-NMS. *Ecological Informatics*. 2023;73:101931. <https://doi.org/10.1016/j.ecoinf.2022.101931>.
- 17) Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. ByteTrack: Multi-object Tracking by Associating Every Detection Box. *Computer Vision ECCV 2022: 17th European Conference*. 2022. https://doi.org/10.1007/978-3-031-20047-2_1.
- 18) Kaur H, Sahambi JS. Vehicle Tracking in Video Using Fractional Feedback Kalman Filter. *IEEE Transactions on Computational Imaging*. 2016;2(4):550–561. <https://doi.org/10.1109/TCI.2016.2600480>.
- 19) Wong TT, Yeh PY. Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*. 2020;32(8):1586–1594. <https://doi.org/10.1109/TKDE.2019.2912815>.