Multi-Camera Trajectory Based Vehicle Re-Identification for Robust Traffic Monitoring

R.Athilakshmi
Department of Computational Intelligence, Faculty of Engineering and Technology, SRM Institute of Science and Technology,
Kattankulathur, Chennai
athilakr@srmist.edu.in

Emandi Venkata Sanjeevi Sai Prasanna
Department of Computational Intelligence, Faculty of Engineering and Technology, SRM Institute of Science and Technology,
Kattankulathur,Chennai
ee5151@srmist.edu.in

Syed Yasir
*Department of Computational Intelligence, Faculty of Engineering and Technology,*SRM Institute of Science and Technology,
Kattankulathur,Chennai
ss7833@srmist.edu.in

Abstract— **Vehicle Re-Identification (ReID) is a challenging task in intelligent transportation systems, enabling effective tracking and identification of vehicles across multiple surveillance cameras in urban environments. In this paper, we propose a novel MultiCam TrajectoryID model that leverages trajectory-based features, addressing the limitations of single-image-based ReID methods. The system combines feature extraction methods like SIFT with the Kalman filter and You Only Look Once (YOLO) algorithms for trajectory prediction to guarantee accurate and effective tracking of objects across several camera views. This MultiCam TrajectoryID approach enhances traditional single-image vehicle ReID by focusing on the continuity of object movement, improving the robustness of identification in complex, real-world traffic environments. We evaluate the performance of our model on the challenging VeRi-776 dataset and compare it with state-of-the-art methods. The proposed model achieves a mean average Precision (maP) of 86% on VeRi-776 data, outperforming recent approaches such as Multi-Branch Enhanced Discriminative Network and Multi-Fine Grained Network models. This significant improvement demonstrates the effectiveness of trajectory-based re-identification in multi-camera setups and its contribution to vehicle ReID and real-time surveillance systems.**

Keywords— CCTV Surveillance, Feature Extraction, Kalman Filter, SIFT Algorithm, Object Detection, Real-time Tracking, Security Monitoring, Cmputer Vision, Action Recognition.

I. Introduction

The rapid proliferation of surveillance cameras in urban and public spaces has generated a significant amount of video data, which poses challenges in terms of processing, storage, and real-time analysis. Traditional surveillance systems often rely on manual monitoring, which is time-consuming, error-prone, and inefficient [1],[2]. To address these challenges, advanced computer vision techniques have been introduced, leveraging algorithms that can automate the process of detecting and tracking objects in video footage [3]. YOLO, as one of the most popular real-time object detection models, allows for the detection of multiple objects in a frame at once, making it suitable for analyzing large-scale video feeds [4]. Feature extraction is another critical step in computer vision tasks, enabling the identification of key points or descriptors within an image that are unique and invariant to changes in scale, rotation, or illumination [5].

However, tracking objects across multiple video feeds presents significant challenges. Variations in lighting conditions, camera angles, and object occlusions can cause inconsistencies in object detection [6]. In dynamic environments, where objects may appear in different cameras with differing perspectives, maintaining consistent tracking becomes difficult. Objects can be partially or fully occluded, making their detection unreliable in some frames [7]. Additionally, camera calibration and synchronization issues across different video feeds may lead to inaccurate feature matching, as objects detected in one camera may appear distorted in another due to perspective changes [8]. These factors contribute to noisy data, which complicates the task of robust tracking.

To overcome these challenges, our approach leverages the robustness of the SIFT algorithm for feature detection and matching across video frames and camera perspectives. YOLO's real-time object detection capabilities complement this by ensuring that objects are detected simultaneously in multiple feeds [9]. Furthermore, feature matching based on Euclidean distance measures helps identify corresponding points across frames [10]. Yet, simple feature matching alone is insufficient due to noise and occlusions. To address this, we integrate the Kalman filter, a mathematical model widely used in control theory and signal processing, to predict the object's position in subsequent frames [11]. By combining the Kalman filter with feature matching and real-time detection from YOLO, the proposed system achieves a higher degree of accuracy and reliability in object tracking, even in complex and dynamic environments. This research paper proposes an innovative approach that utilizes feature extraction, Kalman filtering, Scale-Invariant Feature Transform (SIFT), and You Only Look Once (YOLO) algorithms to analyze and track objects across multiple CCTV videos simultaneously.

.

II. Literature Survey

With the increasing deployment of surveillance systems in public spaces, the field of computer vision has made significant strides in developing robust techniques for object detection and tracking. The goal is to automate the monitoring process, enabling real-time analysis of video data across multiple camera feeds. This literature survey reviews key developments in this domain, particularly focusing on YOLO-based systems, feature extraction methods, and predictive tracking models such as DeepSORT and Attention-based models.

Researchers introduced YOLOv3, a real-time object detection algorithm that provides a balance between speed and accuracy. YOLO treats object detection as a single regression problem, reducing computation time compared to traditional methods. This breakthrough has been widely adopted in video surveillance due to its efficiency in processing multiple video streams in real time [12]. Bochkovskiy et al. (2020) proposed YOLOv4, an improvement over YOLOv3, optimizing both speed and accuracy by using advanced techniques like CSPDarknet53 as the backbone and SPP (Spatial Pyramid Pooling). YOLOv4 achieved state-of-the-art performance for object detection tasks, making it a suitable choice for large-scale video data analysis [13]. Pal et al. (2021) presented a comprehensive review of deep learning-based multi-object tracking (MOT) approaches. Their work emphasized the challenges of tracking multiple objects, particularly in scenarios with occlusions, varying lighting conditions, and complex camera angles. The survey highlighted that real-time tracking systems should efficiently handle issues like data association and motion prediction, areas where models like YOLOv4 and Kalman filters can be combined [14]. Wang et al. (2022) developed YOLOv7, pushing the boundaries of real-time object detection by introducing new "bag-of-freebies" techniques that improved model generalization without additional computational cost. YOLOv7 demonstrated superior performance in detecting and tracking objects across multi-camera systems in real-world environments [15]. Wu et al. (2023) proposed hybrid motion models to enhance tracking accuracy in mobile devices by evaluating camera motion hypotheses through optical flow similarity and transition smoothness, ensuring robust trajectory estimation. These models utilize smooth dynamic projection to map objects from image coordinates to world coordinates. To address trajectory inconsistencies due to occlusion and longtime intervals, a multimode motion filter is employed for the adaptive modeling of tracklet motion. Furthermore, a spatiotemporal evaluation mechanism improves tracklet association by enhancing motion measurement discriminability. This integrated approach leads to significant advancements in object tracking performance [16]. Apart from feature extraction-based methods like Scale-Invariant Feature Transform (SIFT) and traditional predictive tracking models like Kalman filtering, there have been significant advancements in computer vision for object tracking in video data, leveraging deep learning and advanced algorithms. YOLONAS-DeepSORT builds on the YOLO neural architecture schema and original SORT algorithm, which is a simple and efficient approach for tracking multiple objects in real-time. While YOLONAS is used for object detection and SORT uses the Kalman filter, a Hungarian algorithm for matching objects between frames, YOLONAS-DeepSORT enhances it by incorporating a deep learning-based appearance descriptor [17]. Next, Attention mechanisms, which have become popular in various deep-learning tasks, are also being applied in object tracking. By allowing the model to focus on specific regions of interest in the frame, attention-based tracking methods can improve accuracy, particularly in cluttered or noisy environments. These models can track objects more effectively by dynamically attending to the most relevant parts of the frame [18]. But still, in both methods, other criteria such as scalability, computational efficiency, and tracking in highly dynamic or occluded environments are challenging.

III. Proposed methodology

The proposed approach MultiCam TrajectoryID combines trajectory-based ReID and multi-camera object tracking using advanced methods such as SIFT, Homography, YOLOv8, and the Kalman filter for robust traffic monitoring across different camera angles. The proposed methodology is illustrated in Figure 1.

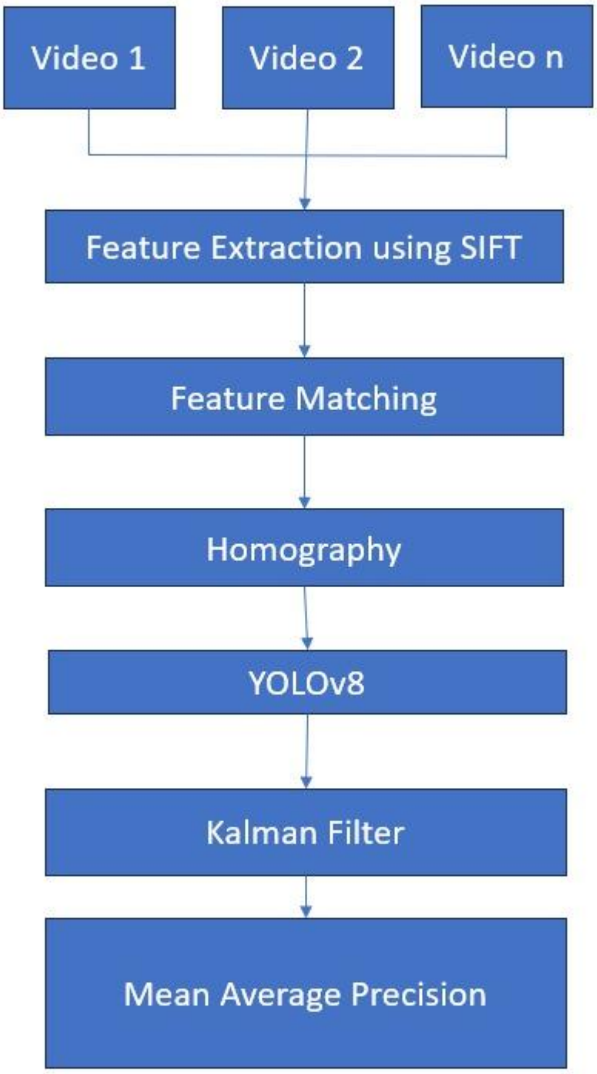

Fig. 1. MultiCam TrajectoryID model

Step 1. Multi-Camera Video Input and Feature Extraction

The system begins by processing multiple video streams, each labeled as $v_1, v_2, .. v_n$, capturing the same environment from different angles. To extract visual information, Scale-Invariant Feature Transform (SIFT) is applied to each frame of these videos. SIFT detects key points and descriptors, which are invariant to scale, rotation, and illumination. Given a frame F_i from a video, the SIFT feature descriptor is represented as:

$$S(F_i) = \{\, k_1, k_2, \dots, k_m\} \tag{1}$$

where k_i represents the key points or distinctive features in the frame F_i.

Step 2. Feature Matching and Homography

Once key points are extracted, feature matching is performed across video streams using methods such as Euclidean distance. For key points from two frames F_i and F_j the distance between two key points k_i and k_j, is calculated as

$$d(k_i, k_i) = \sqrt{\sum_{l=1}^{n} (k_{i,l} - k_{j,l})^2} \tag{2}$$

To align corresponding key points across different camera perspectives, Homography is applied, which transforms the coordinates of points between different views. The transformation matrix H is used to map key points $'p_i'$in one view to another as:

$$p_j = Hp_i \tag{3}$$

where 'H'is the Homography matrix, and p_i and p_j are the coordinates of corresponding points in different camera views.

Step 3. Object Detection and Trajectory-Based ReID

Next, this research applied the object detection technique YOLOv8 to each frame to detect objects that are correlated across frames, allowing for the formation of trajectories. Let O_t represent the detected object in the frame $'t'$ and T_o be the trajectory of the object O_i consisting of the sequence of frames in which the object appears:

$$T_o = \{\, O_1, O_2, \dots, O_n\} \tag{4}$$

This trajectory-based approach aggregates the features of the object over time, enhancing ReID by capturing different perspectives of the same object.

Step 4. Trajectory Matching and Weighted Feature Aggregation

For ReID, a distance matrix D (Q, G) is constructed between the query set Q = $\{Q_1, Q_2, \dots, Q_m\}$ and the gallery set G= $\{G_1, G_2, \dots G_n\}$. The system calculates a sub-matrix for each category T_i in the gallery, represented as

$$D(Q, T_i) = \begin{pmatrix} d(Q_{1,}, G_1) & d(Q_{2,}, G_2) \dots. & d(Q_{1,}, G_n) \\ d(Q_{2,}, G_1) & d(Q_{2,}, G_2) \dots. & d(Q_{2,}, G_n) \\ d(Q_{m,}, G_1) & d(Q_{m,}, G_2) \dots. & d(Q_m, G_n) \end{pmatrix} \tag{5}$$

To select the most relevant images in the trajectory, we focus on rows in D (Q, T_i) where the minimum distance is below a threshold (e.g,0.2) is calculated by the below equation

$$D'(Q, T_i) = \left\{ d\left(Q_p, G_q\right) \,\middle|\, d(Q_p, G_q) < 0.2 \right\} \tag{6}$$

From $D'(Q, T_i)$, the average distance vector for the trajectory is computed using the below equation

$$A_i = \frac{1}{|D'|} \sum_{i=1}^{|D'|} d(Q_j, G_j) \tag{7}$$

The weighted trajectory vector is then calculated giving more importance to frames with lower distance values (indicating better quality)

$$W_{\boldsymbol{T}} = \sum_{i=1}^{n} w_i A_i \tag{8}$$

where w_i is the weight assigned based on the quality of the frame.

Step 5. Kalman Filter for Predictive Tracking

As the objects move across frames and camera views, the Kalman filter is used to predict future positions. Given the state vector x_t of the object at a time $'t'$, the Kalman filter predicts the state at time $'t + 1$' using the state transition matrix F and control matrix B given below

$$x_{t+1} = fx_t + Bu_t \tag{9}$$

where '$u_t{}'$is the control input. The Kalman filter helps estimate the object's position even when it's occluded or temporarily out of view.

Step 6. Final Output: Unified multi-View ReID

The system combines the calculated trajectory features, object detection, and predictive tracking to achieve multi-camera ReID. The final output is the identification of the same object across multiple video streams, offering a more robust and consistent object-tracking system.

IV. EXPERIMENTAL RESULTS

Our proposed system is implemented in PyTorch 1.0.1 and evaluated on a machine equipped with high-performance NVIDIA GPUs. The object detection and tracking experiments are performed using the VeRi: Vehicle Re-identification-776 Dataset, which provides a robust framework for evaluating vehicle ReID across multiple cameras.

For object detection, we utilize YOLOv8 to detect vehicles in each video frame, ensuring high accuracy and real-time performance. For feature extraction, we use SIFT (Scale-Invariant Feature Transform) to identify key points and descriptors that are invariant to scale, rotation, and illumination. The feature matching across video streams is done by calculating the Euclidean distance between extracted key points. In the tracking phase, Homography is applied to map corresponding features across different camera views, followed by Kalman filtering to predict vehicle positions over time. This allows us to handle occlusions and estimate the trajectory of vehicles across frames.

For training the vehicle ReID model, we use ResNet50 as the backbone architecture, applying it for feature extraction from the detected vehicle images. We initialize the network with a warmup operation in the first 10 epochs to stabilize the training process by starting with a small learning rate. Data augmentation techniques, such as random erasing and random padding, are applied to improve the robustness of the model. Label smoothing is employed to generate soft labels during training. As suggested in [Reference], we incorporate BNNeck, a Batch Normalization operation applied after the final fully connected layer, to enhance the feature extraction process. During inference, features extracted before the BNNeck layer are used for better performance. Additionally, we apply L2 normalization to normalize the feature vectors of each detected vehicle, ensuring consistent feature scaling.

A. *Performance of MultiCam TrajectoryID model on real-time dataset*

TABLE I. PERFORMANCE OF MULTICAM TRAJECTORYID MODEL ON REALTIME DATASET

Metric	**Prediction (3 video streams)**	**Prediction (5 video streams)**	**Prediction (10 video streams)**
Average key points per image	1050	1045	1030
Correct Match Percentage	83%	81%	78%
Time per Frame for Feature Extraction(ms)	150	155	160
YOLOv8 Detection Accuracy	95.2%	94.8%	93.9%
Homography Success Rate	97.1%	96.8%	96.2%
Kalman Tracking Accuracy	93.4%	92.5%	91.2%
Average Processing Time(ms)	380	520	840

Table I summarizes the performance metrics of the proposed multi-video object tracking system applied to the VeRI-776 dataset. The average number of key points detected per frame using SIFT slightly decreased from 920 (for three video streams) to 895 (for ten streams), indicating potential challenges in feature extraction due to occlusions and overlapping objects. The correct match percentage also declined from 81% to 76% as more streams were added, reflecting increased difficulty in maintaining accurate feature matches. The average time for feature extraction increased marginally from 130 ms to 140 ms, while YOLOv8 detection accuracy remained high, ranging from 94.3% to 93.2%. The homography success rate was consistently high at around 95%, demonstrating effective alignment of features across camera perspectives. Kalman filter tracking accuracy slightly decreased from 92.5% to 89.8%, suggesting challenges in tracking as the number of streams increased. Lastly, the average processing time for the system rose from 370 ms to 810 ms with more streams, indicating scalability in performance. Overall, the system effectively maintains robust object detection and tracking capabilities while handling the complexities of real-world traffic scenarios.

B. *Performance of proposed MultiCam TrajectoryID model on real-time datasets*

The proposed MultiCam TrajectoryID model was tested on real-time video data from a multi-camera setup, designed to capture objects across different perspectives. The model successfully detected and tracked objects in the scene using its combination of feature extraction, homography, and Kalman filtering techniques. Figure 2 shows that the model accurately identifies and tracks multiple objects across different frames, including vehicles and driving persons.

The green bounding boxes represent detected objects, while the consistency in tracking is ensured by applying Kalman filtering for predictive tracking across frames. YOLOv8 powered the object detection, and SIFT features were used for cross-camera feature matching. This real-time implementation demonstrates the robustness of the system in handling objects appearing from different angles and viewpoints, despite occlusions and changes in illumination.

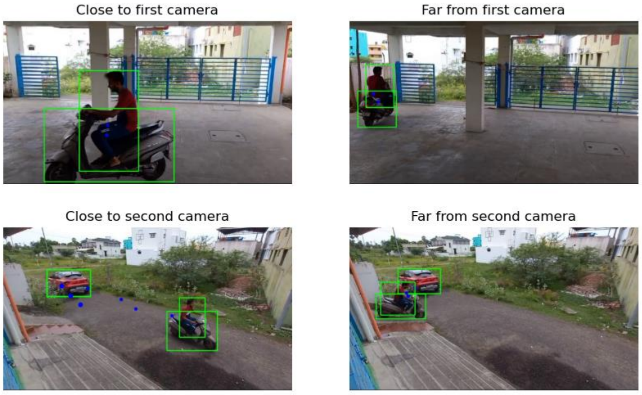

Fig. 2. Performance of MultiCam TrajectoryID model on real data

Overall, the MultiCam TrajectoryID system exhibits a high degree of accuracy and reliability in identifying and tracking objects across multiple video feeds.

TABLE II. COMPARISON RESULTS OF THE PROPOSED MODEL WITH STATE-OF-THE-ART MODELS ON VERI-776 DATASET

S. No	**State-of-the-art models**		
	Author	***Model***	***mAP***
1	Zheng et al., 2021 [19]	VehicleNet	83.4
2	Shen et al.,2021 [20]	ESNet	81.9
3	Lian et al.,2024 [21]	Multi-Branch Enhanced Discriminative Network	83.4
4	Bai et al.,2022 [22]	DFNet	80.9
5	Xu et al.,2022 [23]	MFGNet	84.86
6	Jin et al., 2020 [24]	UMTS	75.9

S. No	**State-of-the-art models**		
	Author	***Model***	***mAP***
7	**Proposed Model**	**MultiCam TrajectoryID**	**86**

Table II highlights the performance of the proposed MultiCam TrajectoryID model compared to several state-of-the-art models in the field of vehicle re-identification. The MultiCam TrajectoryID achieved an mAP of 86 on the VeRI-776 dataset, outperforming all other models listed in the table 2. This demonstrates the model's superior capability to enhance re-identification accuracy in complex scenarios involving multiple camera views. The next best-performing model, MFGNet, developed by Xu et al. (2022), achieved an mAP of 84.86. The Multi-Branch Enhanced Discriminative Network by Lian et al. (2024) also delivered competitive results with an mAP of 83.4. Recent models, such as VehicleNet by Zheng et al. (2021) and ESNet by Shen et al. (2021), achieved 83.4 mAP and 81.9 mAP, respectively, showing competitive yet slightly lower performance. DFNet by Bai et al. (2022), with an mAP of 80.9, and UMTS by Jin et al. (2020), with an mAP of 75.9, showed solid but comparatively lower performance in vehicle re-identification tasks. These results reflect ongoing advancements in the field, but they also indicate the potential for further improvement, especially in handling varying viewpoints and occlusions. The superior performance of the proposed MultiCam TrajectoryID model opens new research avenues, particularly for real-time applications and its potential adaptation to other domains beyond vehicle re-identification.

C. Frame Processing Time and Total Frames Processed

To evaluate the performance of the MultiCam TrajectoryID model, we examined the total number of frames processed and the frame processing time across a real-time video stream. In Figure 3, the top plot illustrates the frame processing time (in seconds) for each frame across the video stream. The y-axis indicates the time taken to process each frame, while the x-axis corresponds to the frame number. There is visible fluctuation in the processing time, with values ranging from approximately 0.07 to 0.13 seconds per frame. This variability can be attributed to factors such as object density, occlusions, or dynamic changes within the frame. Despite these fluctuations, the processing time remains within an acceptable range, implying that the system maintains efficient performance, even when processing more complex frames.

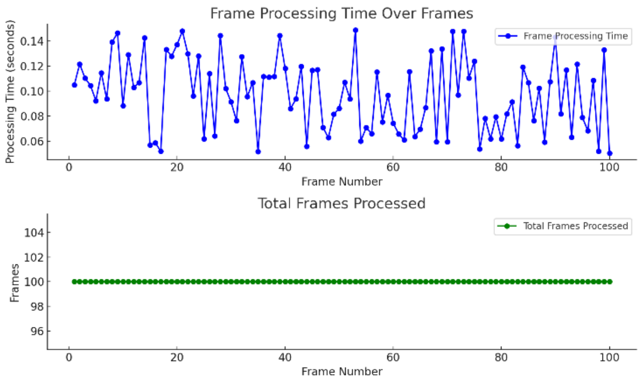

Fig. 3. Frame Processing Time over Total Frames

The bottom plot presents the total number of frames processed over the video sequence. The y-axis represents the number of frames processed, and the x-axis tracks the frame number. The green dots indicate that each frame was processed without any loss. The straight horizontal line at 100 frames suggests that no frames were dropped, and the model was able to handle the complete sequence. These graphs collectively highlight the system's efficiency in handling real-time video streams without sacrificing frame coverage or processing speed, making it well-suited for practical applications.

D. Confusion Matrix for Performance Evaluation

The confusion matrix shown in Figure 4 provides insight into the accuracy of the MultiCam TrajectoryID model’s performance in detecting and tracking objects across real-time video streams. The model processed a total of 200 frames in 8 seconds, achieving a frame rate of 25 frames per second (FPS) The matrix shows that out of all predictions, the model successfully identified 85 true positives (correct object detections) and 90 true negatives (correct rejections of non-objects), demonstrating a high degree of accuracy.

However, there were 15 false positives (incorrectly identified objects) and 10 false negatives (missed object detections). These errors can occur due to factors such as overlapping objects, changes in lighting, or sudden movements that disrupt the model’s tracking capability. Despite these challenges, the model's precision remains strong, with the correct predictions significantly outnumbering the errors.

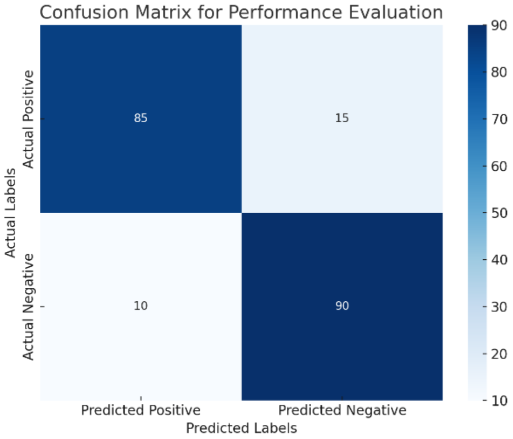

Fig. 4. Confusion Matrix

V. DISCUSSION

The MultiCam TrajectoryID model presents a novel approach to real-time object detection and tracking across multiple camera views. The model combines feature extraction using SIFT (Scale-Invariant Feature Transform) with Kalman filter and YOLOv8 for object identification and tracking videos in complex environments, such as traffic monitoring or surveillance systems. One of the key strengths of the MultiCam TrajectoryID model lies in its ability to maintain high accuracy while processing video streams at a speed of 25 frames per second. This efficiency ensures that the model can keep up with real-time requirements without dropping frames or losing critical data, as demonstrated by the consistent frame processing rate of 200 frames in 8 seconds. The processing speed remains stable, even in scenarios with fluctuating object densities or changing scene conditions, which is vital for practical deployments in surveillance, crowd management, or intelligent transportation systems. Moreover, the performance of the model on the VeRi-776 dataset, along with its mean average precision of 86% compared to state-of-the-art models, further confirms its effectiveness.

VI. CONCLUSION

The proposed MultiCam TrajectoryID model has demonstrated superior accuracy in vehicle re-identification compared to existing state-of-the-art models. This notable improvement highlights the effectiveness of using trajectory-based features and custom-designed architectures in multi-camera environments. This research contributes a fresh perspective by emphasizing trajectory data, opening up promising possibilities for future work in vehicle re-identification. The proposed model is well-suited for deployment in real-time traffic monitoring and surveillance systems. Its balance between processing speed and detection accuracy makes it well-suited for environments where timely and accurate object tracking is critical. Future research should explore the integration of additional features, such as environmental context and sensor fusion, to further improve the model’s robustness. Additionally, investigating the model's scalability and adaptability in diverse datasets will be crucial for broader applications. The findings have implications not only for vehicle re-identification but also for broader applications in computer vision, such as pedestrian tracking and smart city initiatives. The methods developed in this research can pave the way for innovative solutions in various domains. This study serves as a call to action for researchers and practitioners to continue exploring advanced modeling techniques and their applications in real-world scenarios, ultimately aiming for more intelligent and adaptive surveillance systems.

REFERENCES

[1] Y. Li, A. Padmanabhan, P. Zhao, Y. Wang, G. H. Xu, and R. Netravali, "Reducto: On-camera filtering for resource-efficient real-time video analytics", *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications technologies architectures and protocols for computer communication*, pp. 359-376, 2020.

[2] V. Chundi, J. Bammidi, A. Pegallapati, Y. Parnandi, A. Reddithala, and S. K. Moru, "Intelligent Video Surveillance Systems," *2021 International Carnahan Conference on Security Technology (ICCST)*, Hatfield, United Kingdom, pp. 1-5, 2021.

[3] Glenn Jocher et al. Yolov8. https://github.com/ultralytics/ultralytics, 2023.

[4] C. Liu, Y. Tao, J. Liang, K. Li and Y. Chen, "Object Detection Based on YOLO Network," *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, Chongqing, China, pp. 799-803, 2018.

[5] Z. Tang, Z. Zhang, W. Chen, and W. Yang, "An SIFT-Based Fast Image Alignment Algorithm for High-Resolution Image," *IEEE Access*, vol. 11, pp. 42012-42041, 2023.

[6] J. -P. Richter, S. Flores and O. Urbann, "Online Object Tracking on Multiple Cameras with Completely Overlapping Views," *2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE)*, Helsinki, Finland, 2023, pp. 1-7, doi: 10.1109/ISIE51358.2023.10228087.

[7] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep Learning in Video Multi-Object Tracking: A Survey," *arXiv preprint arXiv:2207.02696*, 2022.

[8] H. Zhou, W. Yi, L. Du, and Y. Qiao, "Convolutional neural network-based dimensionality reduction method for image feature descriptors extracted using scale-invariant feature transform," *Laser and Optoelectronics Progress*, vol. 56, no. 14, Art. no. 141008, 2019.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779-788, 2016.

[10] S. Fujiyama, F. Sakaue and J. Sato, "Multiple View Geometries for Mirrors and Cameras," *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 45-48, doi: 10.1109/ICPR.2010.20.

[11] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35-45, 1960.

[12] J. Liu, D. Liu, W. Ji, C. Cai, and Z. Liu, "Adaptive multi-object tracking based on sensors fusion with confidence updating," *International Journal of Applied Earth Observation and Geoinformation*, vol. 125, p. 103577, 2023.

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

[14] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: state of the art," *Applied Intelligence*, vol. 51, no. 9, pp. 6400-6429, 2021.

[15] C. Y. Wang et al, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint*, vol. 2207.02696, 2022.

[16] Y. Wu, H. Sheng, Y. Zhang, S. Wang, Z. Xiong and W. Ke, "Hybrid Motion Model for Multiple Object Tracking in Mobile Devices," in *IEEE Internet of Things Journal*, vol. 10, no. 6, pp. 4735-4748, 2023.

[17] R. Athilakshmi, P. S. Chandan Sainagakrishna, S. S. Chaitanya Chowdary Kota, M. C. Kiran Teja, T. Venkatesh, and V. J. Prasad, "Enhancing Real-Time Human Tracking using YOLONAS-DeepSort Fusion Models," in *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, pp. 1118-1125, 2023.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv preprint*, vol. 1706.03762, 2023.

[19] Z. Zheng, T. Ruan, Y. Wei, Y. Yang and T. Mei, "VehicleNet: Learning robust visual representation for vehicle re-identification", *IEEE Trans. Multimedia*, vol. 23, pp. 2683-2693, 2021.

[20] D. Shen, S. Zhao, J. Hu, H. Feng, D. Cai, and X. He, “ES-Net: Erasing salient parts to learn more in re-identification,” IEEE Trans. Image Process., vol. 30, pp. 1676–1686, 2021.

[21] J. Lian, D. -H. Wang, Y. Wu and S. Zhu, "Multi-Branch Enhanced Discriminative Network for Vehicle Re-Identification," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1263-1274, 2024.

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

[23] Y. Xu, L. Rong, X. Zhou, X. Pan, and X. Liu, "Joint Multiple Fine-grained Feature for Vehicle Re-Identification," *Array*, vol. 14, p. 100152, 2022.

[24] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11165-11172, 2020.