38th International Electric Vehicle Symposium and Exhibition (EVS38)
Göteborg, Sweden, June 15-18, 2025

3D Multi-Object Tracking via Camera-LiDAR Track-Level Association

Seongju Kim[1], Seyoung Jeong[2], Jiwon Im[2], Jeongbin Yoon[2], Sung-Ho Hwang[2]
[1]*Department of Intelligent Robotics, Sungkyunkwan University, Republic of Korea, sduuu29@gmail.com*
[2]*Department of Mechanical Engineering, Sungkyunkwan University, Republic of Korea*

Executive Summary

Accurate perception of the surrounding environment is fundamental for ensuring the safety of autonomous driving systems, as it directly affects behavior and trajectory planning. Although considerable advancements have been made in object detection and tracking using sensors such as cameras and LiDAR, most existing sensor fusion research remains confined to the detection level, limiting the exchange of sequential information. In this paper, we propose a track-level sensor fusion methodology that integrates temporal information by associating object tracks from independent camera-based and LiDAR-based tracking systems. Evaluation on the nuScenes dataset demonstrates that the proposed approach outperforms single-sensor tracking systems, highlighting its effectiveness in improving the robustness and accuracy of object tracking for autonomous driving applications.

Keywords: Autonomous xEV, Intelligent Transportation System for EVs, AI - Artificial Intelligence for EVs

1 Introduction

To ensure the safety of autonomous driving systems, it is essential to accurately perceive the surrounding environment and use this information to plan the vehicle's future behavior and trajectory. In other words, accurately detecting and predicting the surrounding environment is a core task for safe autonomous driving. Various studies have been conducted to improve the accuracy of the perception, and the most commonly used sensors are cameras and LiDAR.

In the field of computer vision, extensive research has been carried out to recognize and track objects from images acquired by cameras. CNN-based object detection models such as YOLO [1] have been widely used due to their fast real-time performance and decent accuracy. More recently, models like DETR [2], which leverage Transformers originally developed for natural language processing, have also been introduced. Similarly, object recognition using LiDAR has seen the application of methodologies from image-based object detection, with notable examples including 3D object detection models such as PointPillars [3].

As object detection technologies for each sensor have advanced, research on fusing data from both sensors has also been actively conducted. Sensor fusion methods that utilize the geometric transformation relationships between the two sensor modalities [4] have been commonly explored. More recently, deep

learning-based sensor fusion approaches, such as TransFusion [5] and BEVFusion [6], have also been proposed and studied. In addition, research on object tracking has also been actively conducted for each type of sensor. In image-based object tracking, a representative algorithm is SORT [7], which combines the Kalman filter and the Hungarian algorithm. In the field of 3D object tracking, various studies such as AB3DMOT [8] have been conducted based on similar principles.
While object tracking technologies have continued to advance, most existing sensor fusion research has focused on the detection level. Therefore, this paper proposes a sensor fusion methodology by adopting an association between tracks, which are estimated from tracking systems. In order to integrate the temporal information from both sensors, the track-level association algorithm was designed. The performance of the proposed method was validated using the nuScenes dataset [9], and it was confirmed that the performance of the proposed system achieves higher performance compared to single-sensor-based tracking systems.

2 Methods

2.1 Overall Architecture

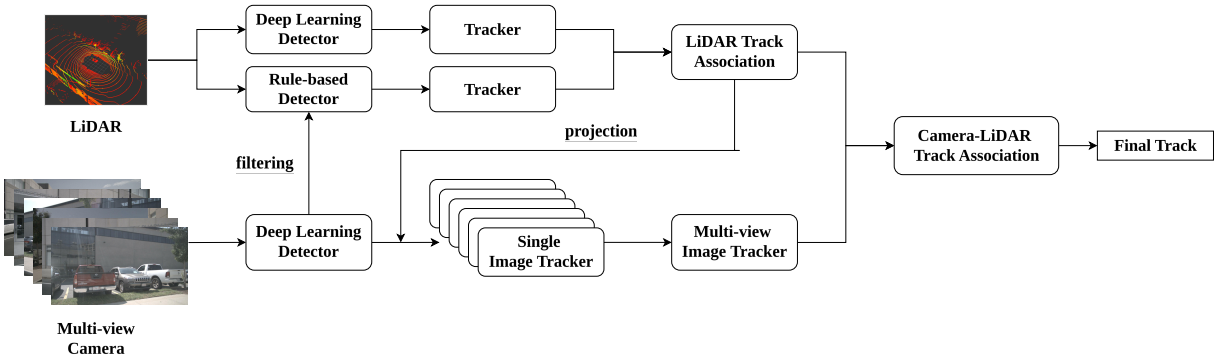

Figure 1: Overall architecture of the proposed system

Figure 1 above describes the overall object detection and tracking system, and the target class of this paper is *Car*. Following [10] and [11], the proposed design consists of multiple associations. However, unlike the previous studies, the system proposed in this paper performs association at the track level. A separate object tracking system for each detection system tracks the objects detected from the sensor data. The tracks estimated by the object tracker are first associated within each modality. After a single modality association, the cross-modality association algorithm filters the final object tracks.
In addition, multiple detection methods were applied to the system. First, two detection systems are applied to the LiDAR data. The deep learning-based detection system detects the position, orientation, and size of objects, while the rule-based detection system identifies object clusters within the road boundary. After passing through separate object tracking systems, the final LiDAR track is estimated through association.
Next, to detect objects from the multi-view images, the deep learning-based detection system was applied to an image batch with a size of the number of cameras. The detected bounding boxes are tracked individually for each image, and the tracks from the multi-view images are then integrated. The final camera tracks obtained are associated with the final LiDAR tracks to estimate the final 3D track.
In addition, a sensor fusion algorithm was designed to facilitate data fusion between two sensors. Since the LiDAR rule-based detection system cannot classify the object's class, false negative detections and non-vehicle class results were filtered by comparing them with the camera detection results. Furthermore, false negative detections in the camera data were addressed by projecting the final LiDAR track onto the image and using it as input for the camera tracking system.

2.2 Object Detection and Tracking Systems

2.2.1 LiDAR Object Detection and Tracking

To detect 3D objects from LiDAR point clouds, both deep learning-based and rule-based object detection algorithms were employed simultaneously. First, for deep learning-based object detection, the PointPillars [3] model was used, which outputs the position, orientation, and size of objects. Next, similar to [12], the rule-based object detection system detects objects' points from LiDAR raw data. It filters point clouds within the road boundary areas defined on map data and then identifies clusters corresponding to objects based on differences in the z-coordinate. As a result, the average x, y, and z coordinates of each cluster can be obtained. Following Figure 2, Figure 3 describes the rule-based detection result and deep learning-based detection result, respectively.

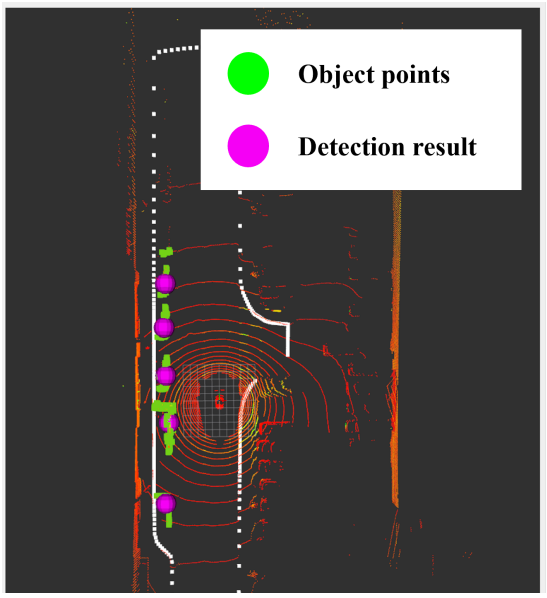

Figure 2: Rule-based detection result

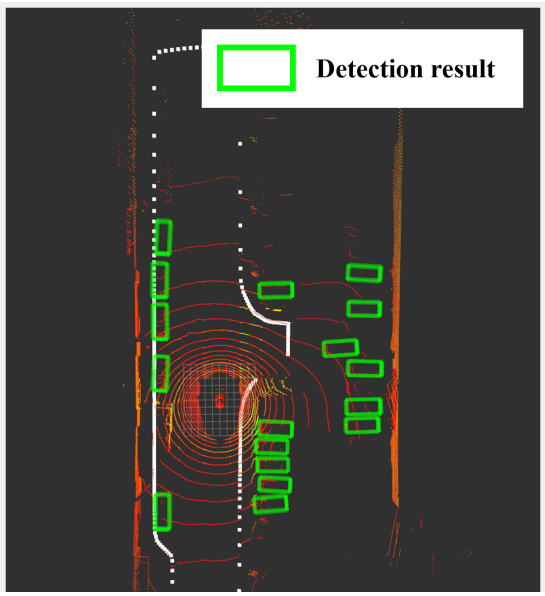

Figure 3: PointPillars detection result

To deal with the differences in the output formats of the two detection systems, a separate Kalman filter was applied to perform object tracking. Since the deep learning-based detection results provide both the position and orientation of objects, a linear Kalman filter with a constant velocity motion model in the global BEV (Bird's Eye View) coordinate system was used. The following describes the state vector of the tracking system of deep learning-based detection and the Kalman filter's prediction step.

$$T_{deep} = [x, y, z, \theta, \dot{x}, \dot{y}, \dot{\theta}]^T \tag{1}$$

$$\begin{bmatrix} x_k \\ y_k \\ z_k \\ \theta_k \\ \dot{x_k} \\ \dot{y_k} \\ \dot{\theta_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & dt & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & dt & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & dt \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ z_{k-1} \\ \theta_{k-1} \\ \dot{x}_{k-1} \\ \dot{y}_{k-1} \\ \dot{\theta}_{k-1} \end{bmatrix} \tag{2}$$

Next, since the rule-based detection results do not provide the object's orientation, a nonlinear term was added to estimate the orientation based on the velocity in the global BEV (Bird's Eye View) coordinate system predicted by the Kalman filter. As a result, an Extended Kalman Filter was designed for the tracking. The following describes the state vector of the tracking system of rule-based detection and the

Kalman filter's nonlinear term for orientation estimation. Since the target class of the system is *Car*, we assumed the size of the object as the average size of the *Car* class of the nuScenes dataset.

$$T_{rule} = [x, y, z, \theta, \dot{x}, \dot{y}, \dot{\theta}]^T \tag{3}$$

$$\theta_k = tan^{-1}(\frac{\dot{y}_{k-1}}{\dot{x}_{k-1}}) + \dot{\theta}_{k-1} \times dt \tag{4}$$

Predicted states of each tracking system are associated with the detection results from the corresponding detection system, using the Hungarian algorithm. For a cost metric of the Hungarian algorithm, an Euclidean distance on the global BEV coordinate system between each track and detection was used. The detection result matched with a specific track is used as the input for the correction step of the Kalman filter for that track, and it updates the track's state.

2.2.2 Camera Object Detection and Tracking

For real-time image object detection, a CNN-based YOLOv5 model was used, which predicts the class and bounding box of each object. To accurately track the detected bounding boxes even at low frame rates, this study designed an object tracking algorithm based on DeepSORT [13]. A deep learning network extracts an appearance vector using a deep neural network and uses the cosine distance between these vectors.

In this study, we incorporated spatial relationships into the cost metric to account for vehicles with similar appearances. Accordingly, in addition to the cosine distance between appearance vectors, the Euclidean distance in the image coordinate system and the ratio of bounding box areas were also included. The network for appearance feature extraction is the same as that of DeepSORT [13]. The following is the architecture of the camera object tracking algorithm. In addition, Kalman filter for the tracking is the same as that of SORT [7].

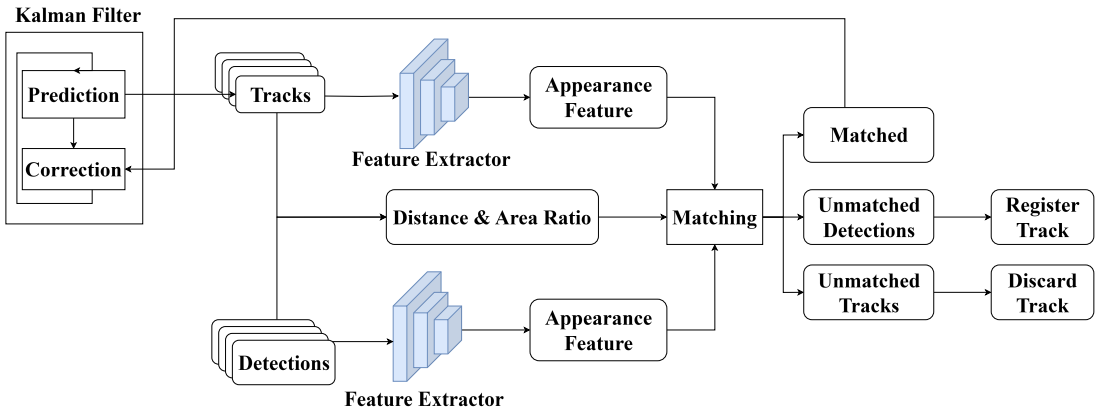

Figure 4: Camera object tracking algorithm

First, the cost metric based on the cosine distance between appearance vectors is defined as follows. $d_{cos}(i, j)$ represents a cosine distance between the appearance vector of the i-th track and the j-th detection.

$$C_1(i, j) = \begin{cases} d_{cos}(i, j), & \text{if } d_{cos} < t_1 \\ \lambda_1 d_{cos}(i, j), & \text{if } d_{cos} \geq t_1 \end{cases} \tag{5}$$

Next, the cost metric based on the Euclidean distance is as follows: $d_{euc}(i, j)$ represents a normalized Euclidean distance between the center of the i-th track and the j-th detection. The normalization term is the Euclidean distance between the farthest corners of the image.

$$C_2(i,j) = \begin{cases} d_{euc}(i,j), & \text{if } d_{euc} < t_2 \\ \lambda_2 d_{euc}(i,j), & \text{if } d_{euc} \geq t_2 \end{cases} \tag{6}$$

Similarly, the following equation describes the cost metric based on the ratio of bounding box areas. $r(i,j)$ represents the box area ratio between the i-th track and the j-th detection.

$$C_3(i,j) = \begin{cases} r(i,j), & \text{if } r < t_3 \\ \lambda_3 r(i,j), & \text{if } r \geq t_3 \end{cases} \tag{7}$$

Lastly, the final cost value between i-th track and j-th detection is defined as follows.

$$C(i,j) = C_1(i,j) + C_2(i,j) + C_3(i,j) \tag{8}$$

With this cost metric, the Hungarian algorithm associates the existing tracks and the detections. After the matching, the update step of the Kalman filter is the same as that of the LiDAR object tracking system.

2.3 Track Association

2.3.1 Multi-View Image Track Association

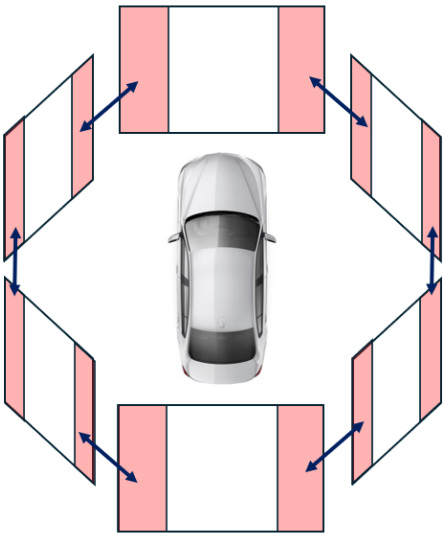

Figure 5: Multi-view image association algorithm

The tracks obtained from different images through single-image object tracking must be integrated and managed within a unified system. To associate the tracks between different images, this paper utilizes the appearance features extracted in the single image object tracking. Specifically, the cosine distance is computed between the appearance features of tracks located at the left and right edges of each image. The association is performed for the tracks obtained from the adjacent images as shown in Figure 5. With this approach, a global ID is assigned to each track, and the system enables matched tracks with the multi-view image association to share the same global ID. The figure below describes the results of multi-view image association. Red bounding boxes indicate tracks matched based on appearance features and thus sharing the same global ID.

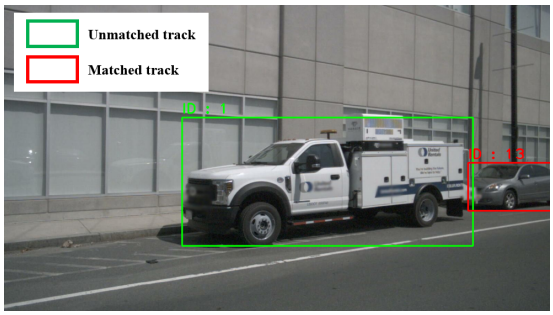

Figure 6: Left front image association result

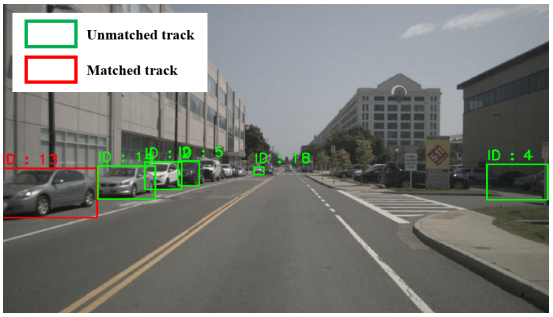

Figure 7: Front image association result

2.3.2 LiDAR Track Association

Tracks obtained from deep learning-based LiDAR object detection and rule-based LiDAR object detection must also be managed within a unified system. To associate LiDAR tracks generated from different filters, this study performs matching based on a cost matrix by calculating the Euclidean distance between tracks from different detection systems. A matching algorithm is then performed using the Hungarian algorithm. Among the matched tracks, the track originating from deep learning-based detection is selected as the final LiDAR track. Unmatched tracks are added to the final LiDAR track set if their Kalman filter states have been updated with observations within the past T_{age} frames. The following describes the results of deep learning-based detection, rule-based detection, and the subsequent track association.

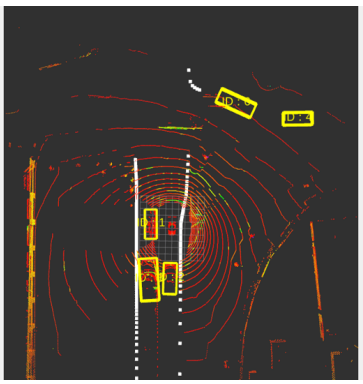

Figure 8: Object tracking result of deep learning-based detections

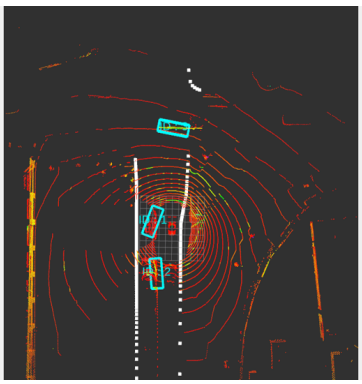

Figure 9: Object tracking result of rule-based detections

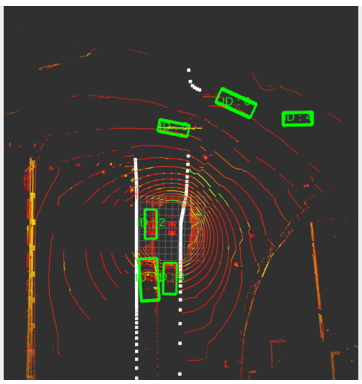

Figure 10: LiDAR track association result

2.3.3 Camera-LiDAR Track Association

Camera and LiDAR tracks, estimated from each modality's object tracking system, exist in different domains. To compare tracks from the different domain, LiDAR tracks are projected onto the image plane for association. Specifically, 2D bounding boxes are generated by projecting the LiDAR tracks onto the image plane, and a cost matrix is constructed based on the negative Intersection over Union (IoU) between the projected LiDAR tracks and the camera tracks. The matching algorithm is then performed using the Hungarian algorithm, and LiDAR tracks that successfully match with camera tracks are selected as final tracks.

LiDAR tracks that do not match with any camera track are used as observations of camera tracker, if a track has been updated with Kalman filter observations within the past T_{age} frames, Using the projected tracks as observations for the image-based object tracking Kalman filter, the system is able to continuously track the corresponding objects by leveraging semantic information available in the image domain.

3 Experiments

3.1 Dataset and Evaluation Metrics

In this study, deep learning model training and object tracking performance evaluation were conducted using the nuScenes dataset. The nuScenes dataset employs a comprehensive set of 3D object tracking metrics to evaluate tracking performance. The primary metric, Average Multi-Object Tracking Accuracy (AMOTA), provides an integrated assessment by averaging MOTAR over various recall thresholds, instead of the traditional metric, the Multi-Object Tracking Accuracy (MOTA) proposed in [14]. In addition, Average Multi-Object Tracking Precision (AMOTP) measures the localization precision of matched objects, averaged across different recall levels. MOTAR is defined as:

$$\text{MOTAR} = max(0, 1 - \frac{\text{FN} + \text{FP} + \text{IDS} - (1 - r) * P}{r * P}) \tag{9}$$

where FN denotes the number of false negatives (missed detections), FP denotes the number of false positives, IDS is the number of identity switches, P represents the total number of ground truth positives for the current class, and r is means recall. Multi-Object Tracking Precision (MOTP), which was also proposed in [14], quantifies the alignment accuracy between predicted and ground truth object positions for matched pairs, and is defined as:

$$\text{MOTP} = \frac{\sum_{i,t} d_{i,t}}{\sum_t TP_t} \tag{10}$$

where $d_{i,t}$ is the position error of track i at time t, between the predicted and ground truth objects for match i at time t, and TP_t indicates the number of matches at t.

3.1.1 Quantitative Analysis

Using these metrics, the proposed method was compared with AB3DMOT[8], which was provided as a baseline algorithm for 3D multi-object tracking in the nuScenes dataset. The performance was evaluated for the *Car* class, and the evaluation range threshold for the class was set as 50m. Following Table 1 is the evaluation result for the nuScenes validation set. As described in the table, our proposed methods achieved higher performance than the baseline method.

Table 1: 3D object tracking results on the nuScenes validation set with range threshold 50m

	AMOTA	AMOTP	IDS
AB3DMOT [8]	0.1339	1.3225	2308
Ours	**0.2294**	**1.2314**	**130**

In addition, the effect of each association system on the overall performance was analyzed. The Table 2 represents the performance evaluation results obtained by progressively adding each association algorithm to the baseline system. The baseline system consists of a single PointPillars [3] model and a single LiDAR object tracker. As shown in the Table 2, object tracking performance was improved through LiDAR track association, and a significant reduction in ID switches was achieved through fusion with the camera by leveraging semantic information from the image domain for association.

Table 2: 3D object tracking results using different association algorithm

	AMOTA	AMOTP	IDS
PointPillars [3] with single tracker	0.1559	1.6933	527
+LiDAR track association	0.1908	1.6438	401
+Camera track association	**0.2294**	**1.2314**	**130**

Lastly, in order to analyze the effect of detection on the tracking performance, we evaluated our system by changing the range threshold. The following Table 3 represents a tracking performance for various range thresholds. As described in the Table 3, as the range threshold decreases, the tracking performance of the system increases. From this result, it was analyzed that the performance of long-range object detection significantly affects object tracking performance.

Table 3: 3D object tracking results with different range threshold

range threshold	AMOTA	AMOTP	IDS
50m	0.2294	1.2314	130
40m	0.3271	1.2153	124
30m	**0.386**	**1.1089**	**107**

3.2 Qualitative Analysis

In addition to the quantitative evaluation, a qualitative analysis was conducted to evaluate the effect of each functional module on object tracking performance.

3.2.1 Appearance-based Camera Object Tracking

First, the influence of appearance features in image-based object tracking was analyzed. The Figure 11, 12 below represents the sequential image frames when the occlusion occurs. It was observed that IoU-based object tracking algorithms failed to perform accurate association, and the track ID was changed due to occlusion, as shown in Figure 13. In contrast, as described in Figure 14, robust association is achieved by adopting an appearance-based cost function, even under occlusion.

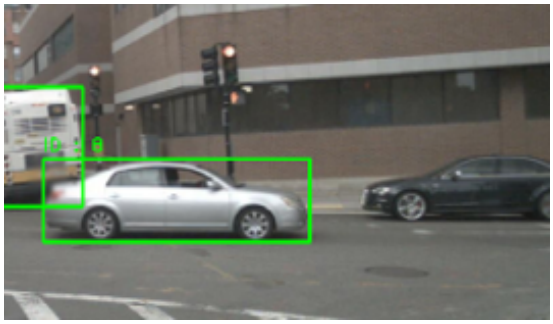

Figure 11: Object tracking result of $(t-1)$-th frame

Figure 12: Object tracking result of t-th frame

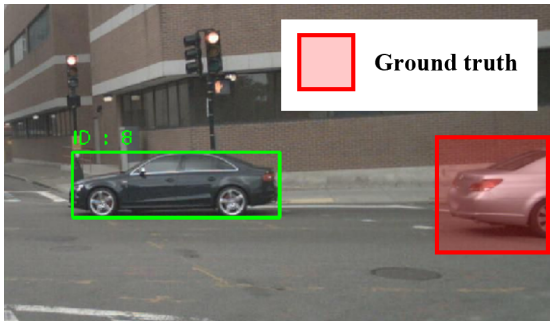

Figure 13: Object tracking result of $(t+1)$-th frame using IoU

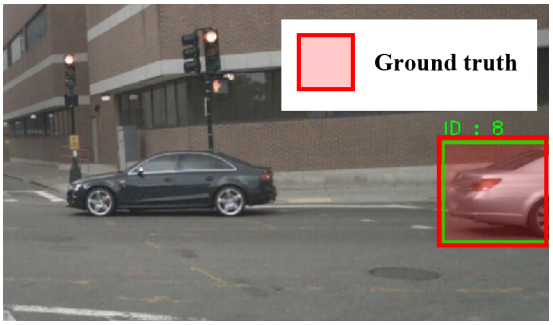

Figure 14: Object tracking result of $(t+1)$-th frame using appearance

3.2.2 Camera-LiDAR Sensor Fusion

Furthermore, an analysis of camera-LiDAR track association was also performed. In the Figure 15, 16 below, it was confirmed that objects missed by the camera due to occlusion or other factors can be compensated for by incorporating LiDAR tracks. In addition, by extracting appearance features from the projected bounding boxes, semantic information from the image domain could be combined to further enhance object tracking performance.

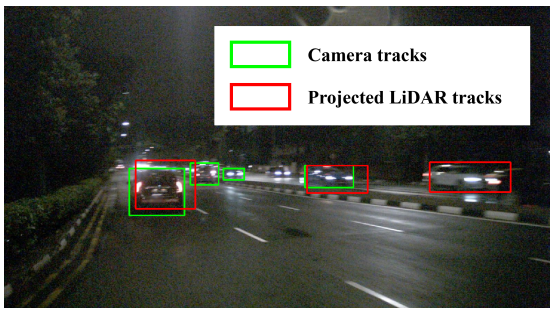

Figure 15: Left front image association result

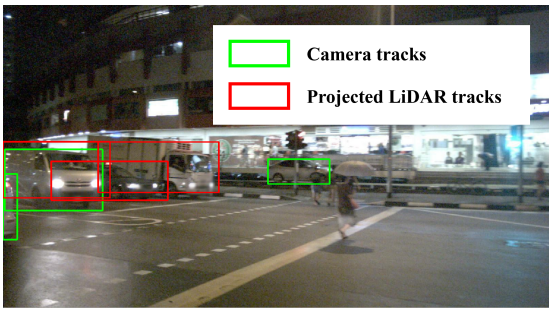

Figure 16: Front image association result

4 Conclusion

This paper proposed a multi-object tracking algorithm based on track-level association. To integrate and manage information obtained from various perception systems, a hierarchical system was designed. First, association is performed among tracks from the same sensor. For camera tracks, the association is conducted based on appearance features, position, and size of the bounding boxes, whereas LiDAR tracks are matched based on their distance in the global coordinate system. Subsequently, a fusion between camera tracks and LiDAR tracks is performed through association on the image plane via projection. The proposed methodology was validated using the nuScenes dataset, and performance evaluation was conducted on the *Car* class within the validation set. The quantitative evaluation demonstrated improvements in AMOTA, AMOTP, and IDS metrics compared to the baseline model. Moreover, the impact of the hierarchical association algorithms and long-range detection on tracking performance was quantitatively analyzed. Furthermore, qualitative analysis confirmed the effectiveness of image-based object tracking using appearance features and track association through camera-LiDAR sensor fusion. However, a limitation of this study is that the evaluation was conducted solely on the *Car* class. Therefore, the generalizability of the proposed system to other object classes has not been verified. Future work will extend the proposed method to a broader range of object classes and the research for the box prediction for the LiDAR object cluster point should be conducted.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00522790).
This work was supported by the Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (RS-2024-00415938, HRD Program for Industrial Innovation)

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[3] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[4] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[5] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[6] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[7] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[8] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 359–10 366.

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[10] X. Wang, C. Fu, Z. Li, Y. Lai, and J. He, "Deepfusionmot: A 3d multi-object tracking framework based on camera-lidar fusion with deep association," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8260–8267, 2022.

[11] A. Kim, A. Ošep, and L. Leal-Taixé, "Eagermot: 3d multi-object tracking via sensor fusion," in *2021 IEEE International conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 315–11 321.

[12] E. Song, S. Jeong, and S.-H. Hwang, "Edge-triggered three-dimensional object detection using a lidar ring," *Sensors*, vol. 24, no. 6, p. 2005, 2024.

[13] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[14] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*, vol. 90, no. 91. Citeseer, 2006.

Presenter Biography

Seongju Kim received a B.S. degree in Mechanical Engineering from Sungkyunkwan University, Suwon, Korea, in 2024. He is currently studying for the M.S. degree in Intelligent Robotics at Sungkyunkwan University. His research interests include sensor fusion, object detection, object tracking, and object trajectory prediction.

Seyoung Jeong received a B.S. degree in Mechanical Engineering from Sungkyunkwan University, Suwon, Korea, in 2021. He is currently studying for the Ph.D. degree in Mechanical Engineering at Sungkyunkwan University. His research interests include localization, sensor fusion, SLAM and autonomous vehicles.

Jiwon Im received a B.S. degree in Mechanical Engineering from Sungkyunkwan University, Suwon, Korea, in 2024. He is currently studying for the M.S. degree in Mechanical Engineering at Sungkyunkwan University. His research interests include sensor fusion, object detection, collision detection and autonomous vehicles.

Jeongbin Yoon received a B.S. degree in Electronic and Electrical Engineering from Sungkyunkwan University, Suwon, Korea, in 2024. She is currently studying for the M.S. degree in Mechanical Engineering at Sungkyunkwan University. Her research interests include sensor fusion, object detection, and autonomous vehicles.

Sung-Ho Hwang received his B.S., M.S., and Ph.D. degrees in Mechanical Design and Production Engineering from Seoul National University, Seoul, Korea, in 1988, 1990, and 1997, respectively. From 1992 to 2002, he worked as a Senior Researcher at the Korea Institute of Industrial Technology in Cheonan. Since 2002, he has been a Professor in the School of Mechanical Engineering at Sungkyunkwan University, Suwon, Korea. He has authored two books, published more than 100 articles, and holds over twenty patents. His research focuses on automotive mechatronics systems of electrified vehicles and control strategies for autonomous vehicles. Prof. Hwang served as Editor-in-Chief of the Journal of Drive and Control from 2012 to 2016 and is currently serving as President of KSAE (The Korean Society of Automotive Engineers) in 2025.