Politecnico di Torino

Department of
Electronics and Telecommunications
MSc. Mechatronic Engineering

01SQHOV - Technologies for Autonomous Vehicles
Prof. Violante

Scalable Multi-Scenario Sensor Fusion for AV

Project Report

Liu Jingting - s336945

Academic Year 2024/2025

1 Introduction

This report presents a practical implementation of sensor fusion for autonomous driving perception using the MAN TruckScenes dataset, combining LiDAR and Radar data via automated spatial alignment and visualization scripts. Results show that the fusion pipeline improves detection robustness across challenging scenarios, supporting the case for multimodal integration in real-world applications. For all details, please refer to https://github.com/LiuJingting0201/AV4.

2 Methodology

2.1 Overall Pipeline and Output Structure

The workflow consists of sample selection, data preprocessing, sensor fusion, visualization, and batch statistical analysis. All outputs are organized by scenario and sample in a standardized directory structure, supporting scalable experiments and reproducibility.

2.2 Scenario Coverage and Selection

Representative scenarios are identified by parsing scene metadata and descriptions. A script filters for target tags (e.g., "snow_city", "night_highway"), and the selected sample tokens are saved to a configuration file. This file is then used to guide downstream batch processing, ensuring coverage of diverse and challenging conditions.
All outputs are organized into structured directories by sample and scenario to facilitate efficient analysis.

2.3 Preprocessing and Sensor Fusion

Sensor fusion is implemented by spatially aligning and merging LiDAR and Radar point clouds using calibration data. The fused data, along with object annotations, are visualized for qualitative and quantitative comparison. Point cloud data are saved in multiple formats (TXT, CSV, PLY), and for each object, the number of LiDAR, Radar, and fused points within its bounding box is calculated. These statistics support recall and density analysis, providing a quantitative evaluation of the fusion effect. The entire workflow is automated from data selection to statistical analysis.

2.3.1 Coordinate Transformation

All sensor point clouds are transformed to the common ego-vehicle coordinate frame using calibration and pose information. The transformation for each point $\mathbf{p}$ can be formulated as

$$\mathbf{p}_{ego} = \mathbf{T}_{global \to ego} \cdot \mathbf{T}_{sensor \to global} \cdot \mathbf{p}_{sensor}, \tag{1}$$

where $\mathbf{T}_{sensor \to global}$ is constructed from the sensor's extrinsic calibration and pose, $\mathbf{T}_{global \to ego}$ is the inverse ego pose matrix, and $\mathbf{p}_{sensor}$ is the point in the sensor frame. This transformation is implemented in the `_transform_points_to_ego()` method, ensuring that all LiDAR and Radar points are spatially aligned for fusion and visualization.

2.3.2 Sensor Fusion

After all point clouds are aligned to the ego frame, sensor fusion is achieved by concatenating the transformed LiDAR and Radar points:

$$\mathbf{P}_{fusion} = \{\mathbf{P}_{LiDAR}, \mathbf{P}_{Radar}\}, \tag{2}$$

where $\mathbf{P}_{LiDAR}$ and $\mathbf{P}_{Radar}$ are the sets of transformed points from all LiDAR and Radar sensors, respectively. In code, this is realized by stacking the corresponding numpy arrays:

```
merged_lidar = np.vstack(all_lidar_points)
merged_radar = np.vstack(all_radar_points)
```

and then merging as needed. The final fused result is used for subsequent statistics and visualization.

2.3.3 Annotation Transformation

Three-dimensional bounding box annotations are also transformed to the ego frame. For each box, both the center position and orientation are converted. The center transformation is given by

$$\mathbf{c}_{ego} = \mathbf{T}_{global \to ego} \cdot \begin{bmatrix} \mathbf{c}_{global} \\ 1 \end{bmatrix}, \tag{3}$$

where $\mathbf{c}_{global}$ is the box center in the global frame. The orientation, originally described by a quaternion in the global frame, is rotated by the inverse of the ego pose rotation matrix. This ensures consistent spatial comparison between annotations and fused points.

2.4 Visualization and Analysis

For each sample, multi-view geometric visualizations are generated, including side, top, and front perspectives of the fused point clouds. Points are colored by distance, and 3D bounding boxes are overlaid to support assessment of coverage and spatial alignment. Camera images and interactive 3D scenes are also produced for qualitative interpretation. In addition, key statistical results are visualized as summary plots, including recall rates by object category, distance range, and scenario. These bar charts provide a quantitative evaluation of detection performance for each modality and for the fusion, facilitating objective comparison under various conditions. All visual outputs are automatically generated and organized for efficient analysis.

2.4.1 Statistical Analysis

For each object bounding box, the number of points inside the box is counted for each modality and their fusion. Given a box with center $\mathbf{c}$, size (l, w, h), and rotation $\mathbf{R}$, a point $\mathbf{p}$ is inside the box if

$$\left|(\mathbf{R}^{-1}(\mathbf{p}-\mathbf{c}))_x\right| < \frac{l}{2}, \quad \left|(\mathbf{R}^{-1}(\mathbf{p}-\mathbf{c}))_y\right| < \frac{w}{2}, \quad \left|(\mathbf{R}^{-1}(\mathbf{p}-\mathbf{c}))_z\right| < \frac{h}{2} \tag{4}$$

This point-in-box logic is implemented in `fusion_stats.py`, using an axis-aligned or oriented bounding box test as appropriate. Recall for each modality is then calculated as the fraction of boxes with at least one point detected:

$$\text{Recall} = \frac{\text{Boxes with } N_{pts} > 0}{\text{Total boxes}} \tag{5}$$

This ensures that recall is defined as the proportion of ground truth boxes for which at least one point has been detected.

2.4.2 Batch Processing and Output Organization

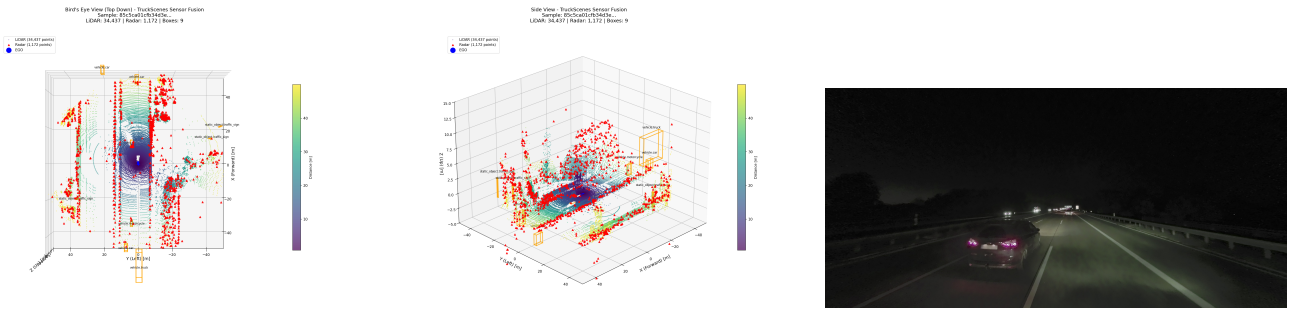
All modules support batch processing by looping over a list of selected sample tokens (from `selected_samples.json`). Results are saved in structured directories with standardized naming conventions for all generated files, including visualizations, point clouds, and statistical summaries. This design enables scalable, reproducible experiments and efficient comparative analysis.

3 Experiments and Results

A comprehensive evaluation has been conducted using multiple representative scenarios from the MAN TruckScenes dataset. Performance metrics for LiDAR, Radar, and their fusion have been compared quantitatively across different object categories, distance ranges, and environmental conditions.

3.1 Visualization Examples

A total of nine scenarios were selected for evaluation, with multiple frames analyzed for each. For clarity and brevity, only two representative scenes—night highway and terminal area—are illustrated here. For each scenario, three images are provided side by side: the bird's eye view (BEV) of the fused point cloud, the 3D side view, and the synchronized front-left camera image simulating the driver's perspective. These visualizations highlight the spatial alignment and coverage achieved by sensor fusion.

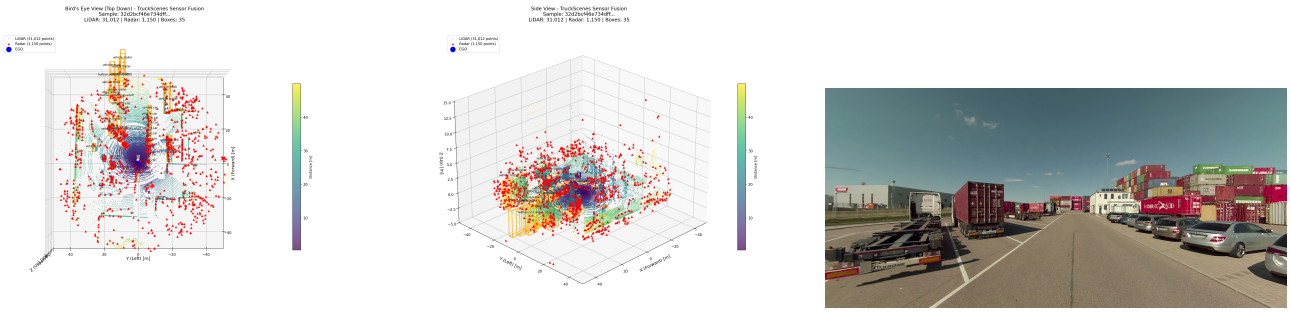


(a) Bird's Eye View (b) 3D Side View (c) Front-left camera (driver)

Figure 1: Visualization examples in the night highway scenario

In the night highway scenario, the fused point cloud achieves clear spatial coverage of distant vehicles and distant objects that are only sparsely represented or even missed by individual LiDAR or Radar. Occluded and low-reflectivity targets are reliably detected in the fusion result, demonstrating the necessity of multi-modal integration for night-time

perception. The driver's view confirms the presence of challenging, low-visibility targets which are successfully captured by sensor fusion. For interactive 3D inspection, see the supplementary file: night_highway_interactive_3d.html



(a) Bird's Eye View (b) 3D Side View (c) Front-left camera (driver)

Figure 2: Visualization examples in the terminal area scenario

In the terminal area scenario, dense occlusions and overlapping objects often lead to fragmented or missing detections when using a single sensor. Sensor fusion produces a more continuous and complete point cloud, enabling the restoration of partially occluded vehicles and static obstacles. The visualization shows that fusion outputs more coherent and spatially aligned bounding boxes compared to individual modalities. For interactive 3D inspection, see the supplementary file: terminal_area_interactive_3d.html

3.2 Scenario-Based Performance

Detection recall for each sensor modality and the fused result has been assessed under various driving scenes. As shown in Figure 3, sensor fusion consistently achieves higher recall rates than any single modality, particularly in challenging conditions such as nighttime or adverse weather.

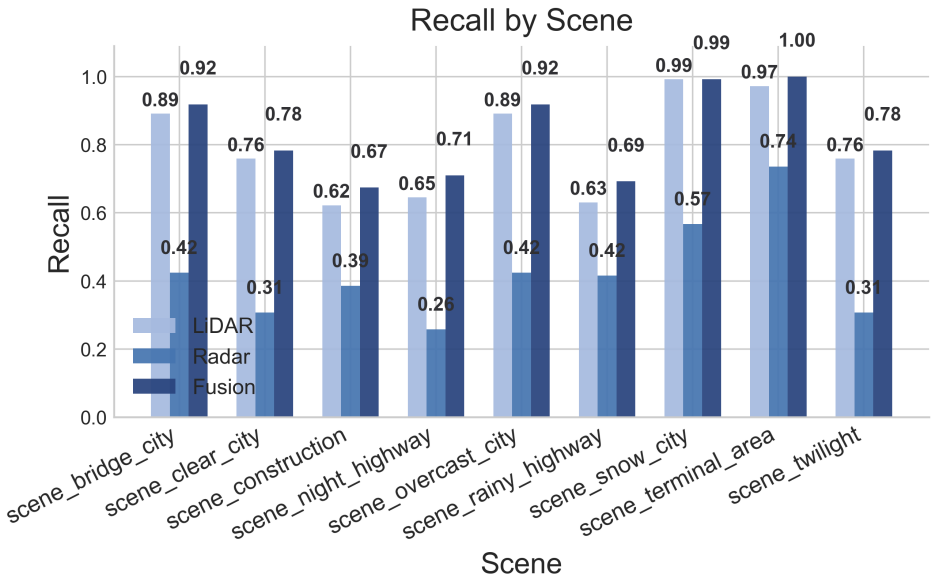


Figure 3: Detection recall by scenario.

3.3 Category-Wise Analysis

Recall by object category is visualized in Figure 4. Superior performance from sensor fusion is observed for most categories, with significant gains for vulnerable road users (e.g., pedestrians, cyclists) and certain vehicle types. This demonstrates the complementary advantages of multi-sensor integration.

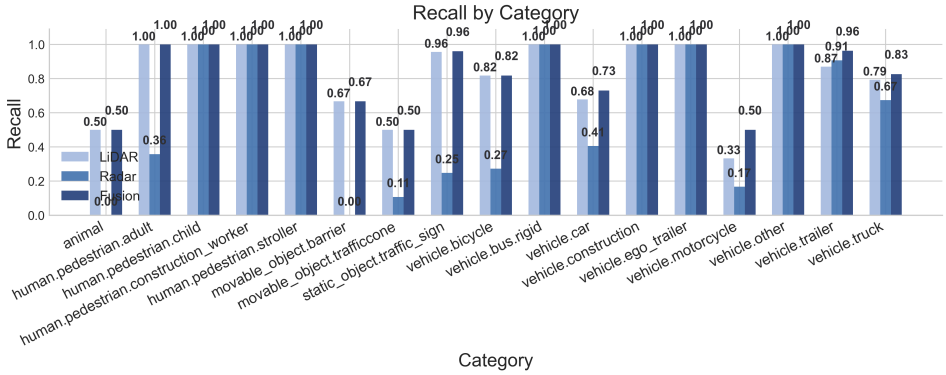


Figure 4: Recall by category.

3.4 Performance Across Distance Ranges

Figure 5 presents recall as a function of object distance from the ego vehicle. It is evident that LiDAR and Radar alone exhibit a decline in detection recall at longer distances, whereas sensor fusion maintains higher recall across all distance bins. This highlights the enhanced robustness provided by multi-modal perception.

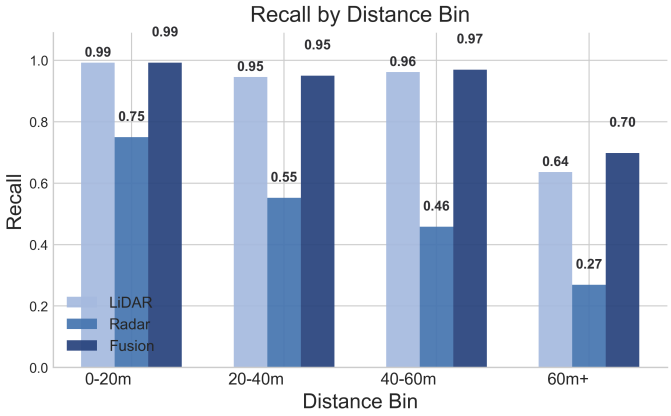


Figure 5: Recall by distance range.

3.5 Summary of Findings

Across all tested scenarios, categories, and ranges, sensor fusion is shown to provide more complete and robust detection coverage compared to any single modality. The automated pipeline enables reproducible, large-scale analysis and objective performance comparison under real-world conditions.

4 Conclusion and Future Work

In this work, a fully automated multi-sensor fusion pipeline has been developed and applied to large-scale autonomous driving perception analysis. LiDAR and Radar point clouds, originally recorded in their respective sensor frames, are precisely transformed into the ego-vehicle coordinate system through a series of calibration and pose operations. Three-dimensional object annotations are also converted from the global frame to the ego frame, ensuring spatial consistency for fusion, visualization, and quantitative analysis.

The implemented workflow supports batch processing and generates structured outputs, including recall statistics and multi-perspective visualizations, for nine representative scenarios. Experimental results demonstrate that sensor fusion significantly improves detection coverage and spatial alignment. This pipeline establishes a scalable and reproducible foundation for comprehensive evaluation of perception systems in real-world settings.

Potential future directions may include the integration of camera data and advanced deep learning-based fusion approaches, such as Transfuser, to further enhance perception robustness and enable end-to-end multi-modal learning.

References

- MAN TruckScenes Dataset and Devkit. https://github.com/TUMFTM/truckscenes-devkit
- ManTruck Official Documentation: https://brandportal.man/d/QSf8mPdU5Hgj