DECISION-LEVEL FUSION OF YOLOV8 AND POINTPILLARS: INITIAL FINDINGS

IVAN VRSALOVIĆ,[1,2] KRISTIJAN LENAC[2]

[1] University of Rijeka, Faculty of Informatics and Digital Technologies, Rijeka, Croatia
ivan.vrsalovic@uniri.hr, klenac@riteh.hr
[2] University of Rijeka, Faculty of Engineering, Rijeka, Croatia
ivan.vrsalovic@uniri.hr, klenac@riteh.hr

This paper presents a multi-modal perception system tailored for autonomous driving and safety monitoring in public parking environments, utilizing a dual-stage decision-level fusion of YOLOv8-seg and PointPillars. The architecture ensures precise occupancy monitoring and safety by integrating 2D instance segmentation with 3D LiDAR point clouds. A specialized decision fusion rngine features a distance-based matching phase and a rescue phase to maintain detections during sensor occlusions. By implementing a 2.5m Euclidean threshold and a high-confidence YOLO override mechanism, the system effectively compensates for LiDAR sparsity. Experimental results on the KITTI dataset demonstrate significant reliability gains, notably increasing the F1-score for pedestrians by 10.11% and cars by 6.99%. These findings prove that the synergy of visual masks and geometric data provides a robust solution for real-time monitoring of vehicles and vulnerable road users (pedestrians and cyclists) in automated parking environments.

DOI
https://doi.org/10.18690/um.feri.4.2026.7

ISBN
978-961-299-111-1

Keywords:
sensor fusion,
LiDAR,
YOLOv8-seg,
PointPillars,
KITTI dataset

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.feri.4.2026.7

ISBN
978-961-299-111-1

ZDRUŽEVANJE MODELOV YOLOV8 IN POINTPILLARS NA RAVNI ODLOČANJA: ZAČETNE UGOTOVITVE

IVAN VRSALOVIĆ,[1,2] KRISTIJAN LENAC[2]

[1] Univerza v Reki, Fakulteta za informatiko in digitalno tehnologijo, Reka, Hrvaška
ivan.vrsalovic@uniri.hr
[2] Univerza v Reki, Tehniška fakulteta, Reka, Hrvaška
ivan.vrsalovic@uniri.hr, klenac@riteh.hr

Ključne besede:
senzorska fuzija,
LiDAR,
YOLOv8-seg,
PointPillars,
nabor podatkov KITTI

Ta prispevek predstavlja multimodalni zaznavni sistem, prilagojen za avtonomno vožnjo in varnostni nadzor v okoljih javnih parkirišč, ki uporablja dvostopenjsko združevanje modelov YOLOv8-seg in PointPillars na ravni odločanja. Arhitektura zagotavlja natančno spremljanje zasedenosti in varnosti z integracijo 2D instančne segmentacije z oblaki 3D točk 3D LiDAR. Specializiran pogon za fuzijo odločitev (decision fusion engine) vključuje fazo umerjanja na podlagi razdalje in fazo reševanja (rescue phase) za ohranjanje zaznav med zakritji senzorjev. Z implementacijo 2,5-metrskega evklidskega praga in mehanizma za preglasitev z visoko stopnjo zaupanja YOLO (YOLO Override), sistem učinkovito kompenzira redkost podatkov LiDAR. Eksperimentalni rezultati na naboru podatkov KITTI izkazujejo znatno izboljšanje zanesljivosti, zlasti s povečanjem ocene F1 za pešce za 10,11 % in za avtomobile za 6,99 %. Te ugotovitve potrjujejo, da sinergija vizualnih mask in geometrijskih podatkov zagotavlja robustno rešitev za sprotno spremljanje vozil in ranljivih udeležencev v prometu (pešcev in kolesarjev) v okoljih avtomatiziranih parkirišč.

University of Maribor Press

1 Introduction

Reliable object detection in dynamic environments remains a significant challenge for autonomous systems. While single-sensor approaches have made considerable progress, they are inherently limited by their physical properties. RGB cameras provide high-resolution semantic data but fail in precise depth estimation and low-light conditions. Conversely, LiDAR sensors provide accurate 3D spatial geometry but suffer from point cloud sparsity, particularly at long ranges or when detecting objects with low reflectivity.

This paper proposes a lightweight, yet robust multi-modal fusion framework designed to bridge these gaps at the decision level. By integrating YOLOv8-seg instance masks with PointPillars 3D proposals, the system achieves a higher level of spatial consistency than standalone models. The primary contribution is the development of a decision fusion engine, which introduces a rescue phase and a high-confidence override mechanism. These features allow the system to maintain detection integrity even when one sensor output is degraded or missing.

While this architecture is applicable to various robotics domains, it is particularly effective in autonomous driving and safety monitoring in parking environments. The system must precisely distinguish between closely parked vehicles and detect vulnerable road users (VRUs), such as pedestrians and cyclists, who are often occluded. Experimental results on the KITTI dataset show that our fusion strategy improves the **F1-score** for pedestrians by **10.11%** and for cars by **6.99%**, proving that decision-level synergy significantly enhances perception reliability.

2 Related works

The field of real-time 3D object detection has shifted significantly toward multi-modal integration to overcome the inherent limitations of single-sensor systems. A cornerstone of modern LiDAR-based detection is the PointPillars architecture(Yao et al., 2025), which revolutionized the domain by voxelizing point clouds into vertical columns, or "pillars". This approach allows the system to utilize efficient 2D convolutional layers for feature extraction, making it highly suitable for real-time applications on edge hardware (Zhang et al., 2024). However, a primary challenge remains the "sparsity problem"; while PointPillars is highly effective for large objects

like cars, its performance often degrades when detecting humans (pedestrians) and cyclists, whose small cross-sections and lower reflectivity result in sparse LiDAR returns (Fei et al., 2020).

To mitigate these issues, recent research emphasizes late-fusion strategies that combine PointPillars with camera-based detectors such as YOLO (You Only Look Once) (Ngo et al., 2025). In a late-fusion or decision-level framework, the LiDAR and camera subsystems process data independently, with their high-level outputs merged in a final stage to ensure system redundancy (Pang et al., 2020). Recent evaluations on the KITTI dataset (Liao et al., 2022) demonstrate that integrating YOLOv8 segmentation masks provides (Malić et al., 2025) critical visual context that "boosts" the detection of human targets (Mitiu, n.d.). By applying a Probabilistic OR operation, the system can combine two relatively weak signals, such as a low-confidence LiDAR cluster and a marginal camera detection to successfully cross the detection threshold (Vora et al., 2020).

Furthermore, the development of robust override mechanisms has become central to maintaining safety in dynamic environments (Soyyigit et al., 2024). Innovations such as the Geometric Fallback mechanism allow the system to remain functional even when LiDAR data is missing or too sparse to project onto a camera mask (Ngo et al., 2025). In such cases, if the YOLO detection confidence is high, the system generates a "Pseudo-3D" bounding box based on camera priors and projection matrices. This ensures that the system maintains a high recall rate for humans and other vulnerable road users, even during sensor-specific failures or extreme sparsity (Mitiu, n.d.). Collectively, these works demonstrate that the synergy between Lidar and Camera fusion of data is essential for creating perception (Vora et al., 2020) pipelines that are both computationally efficient and resilient to environmental noise.

3 Methodology

The perception system is designed as a **late fusion** (He et al., 2023) architecture. This strategy processes data from the camera and LiDAR sensors through independent pipelines before reconciling their outputs at the decision level. This approach ensures high modularity and system reliability; if one sensor provides

sparse or degraded data, the high-level semantic or geometric information from the other sensor can maintain the detection integrity.

3.1 Spatial Synchronization and Calibration

The foundation of the fusion process is the spatial alignment between two distinct coordinate systems. Using a calibration module, the system establishes a geometric link between the 3D LiDAR space and the 2D image pixels. By applying a transformation matrix derived from calibration files, 3D points are projected onto the 2D image plane. This synchronization allows the system to accurately determine which 3D point clusters correspond to specific objects visible in the camera's field of view.

3.2 Dual-Stream Processing

The system utilizes two specialized detection streams that run in parallel:

- **The Semantic Stream**: This stream employs a YOLOv8-seg model to perform instance segmentation. It identifies object classes (such as Cars, Pedestrians, and Cyclists) and generates precise pixel-level masks. These masks are used to isolate specific LiDAR points, effectively filtering out environmental noise like the road surface or background buildings.
- **The geometric stream**: This stream processes 3D proposals generated by the PointPillars model. It provides the system with essential spatial data, including the precise 3D coordinates, object dimensions (length, width, height), and initial orientation.

3.3 The Decision Fusion Engine

The core of the methodology is the integration logic that merges the two streams. The engine calculates the **Euclidean distance** between the 3D centers of the camera-derived objects and the LiDAR proposals. If the centers are within a defined proximity threshold (2.5 meters), the detections are matched and combined.

To account for real-world sensor limitations, the engine follows two critical safety protocols:

- **The rescue phase:** In conditions where the camera may fail to see an object due to poor lighting or occlusion, the engine rescues any high-confidence LiDAR detection to ensure the object is not lost.
- **The high-confidence override:** Conversely, if the LiDAR fails to return a valid proposal due to reflective surfaces or low point density, the engine allows a high-certainty camera detection to override the result and generate a 3D box based on visual data.

3.4 Spatial Refinement and Orientation

Following the fusion of the two streams, the system performs a final geometric refinement. The point cloud associated with each fused object is processed using a clustering algorithm (DBSCAN) (Wang et al., 2019) to remove stray points that do not belong to the physical structure of the target.

To determine the final heading of the object, the system applies principal component analysis (PCA) (Shlens, 2014). By calculating the principal axes of the filtered point cluster, PCA determines the object's yaw (rotation). This mathematical refinement ensures that the final 3D bounding box is accurately aligned with the actual physical orientation of the detected vehicle or pedestrian.

4 Experimental Results and Discussion

The proposed fusion framework was evaluated on the KITTI dataset to assess its effectiveness in improving perception reliability for autonomous driving and safety monitoring in public parking environments. Performance was compared against the standalone PointPillars LiDAR baseline to quantify the benefits of the decision-level fusion strategy under realistic urban parking scenarios.

4.1 Quantitative Analysis

Evaluation focused on the three primary traffic participants: car, pedestrian, and cyclist. Precision, recall, F1-score, and mean intersection over union (mIoU) were used as the main performance indicators (Table 1). The best-performing

configuration was obtained using the probabilistic union fusion with relaxed dual-detection tolerance, which maximized gains for vulnerable road users while maintaining strong vehicle detection accuracy.

Table 1: Experiment results

Method	*Category*	*Precision*	*Recall*	*F1-Score*	*mIoU*
PointPillars (baseline)	Car	0.5953	0.7669	0.6703	0.8251
	Pedestrian	0.2981	0.6098	0.4004	0.6744
	Cyclist	0.5398	0.7376	0.6234	0.7311
Proposed fusion system	Car	0.7017	0.7572	0.7284	0.8255
	Pedestrian	0.4728	0.5596	0.5126	0.6765
	Cyclist	0.5983	0.5144	0.5532	0.7314

The fusion strategy significantly improved detection performance for Cars and Pedestrians. The F1-score for Cars increased from 0.6703 to 0.7284 (+0.0581), while Pedestrian detection showed the largest gain, rising from 0.4004 to 0.5126 (+0.1122). However, these improvements come with a slight trade-off between precision and recall across certain classes. For Cars, precision increased notably (0.5953 → 0.7017) while recall experienced a minor decrease (0.7669 → 0.7572), indicating a more conservative but more reliable detection behaviour. Similarly, for Pedestrians, the fusion system substantially improved precision (0.2981 → 0.4728) at the cost of a modest recall reduction (0.6098 → 0.5596), reflecting fewer false positives but slightly more missed detections.

This precision-recall trade-off suggests that the fusion framework prioritizes detection reliability and spatial consistency over exhaustive recall, which is particularly desirable in safety-critical parking scenarios where false alarms can negatively affect system trust and decision making.

4.2 Impact of Fusion Logic

The substantial improvement in pedestrian detection is primarily attributed to the rescue phase in the decision fusion engine. In parking scenarios, pedestrians frequently appear in shadows, behind vehicles, or partially occluded, which reduces camera-only confidence. However, their vertical structure remains evident in LiDAR point clouds. By preserving these LiDAR-originated detections when camera confidence drops, the system effectively reduces false negatives and improves recall for vulnerable road users.

For vehicle detection, the YOLO override mechanism played a critical role. In distant or sparsely sampled regions, LiDAR may fail to generate sufficiently dense 3D proposals. High confidence 2D masks from the camera were therefore allowed to override missing LiDAR detections, preventing loss of relevant objects and increasing overall precision. This complementary behaviour between modalities demonstrates the advantage of decision-level fusion over isolated sensor pipelines.

The cyclist category exhibited a slight decrease in F1-score compared to the LiDAR-only baseline. This is mainly due to the higher variability in cyclist geometry and motion, which occasionally leads to mismatches during the association stage. Nevertheless, the fusion system maintained comparable mIoU values, indicating that spatial localization remained consistent even when classification confidence fluctuated.

4.3 Qualitative Evaluation

Qualitative inspection of the fused outputs showed that integrating DBSCAN clustering with PCA-based orientation refinement produced more accurately aligned 3D bounding boxes while preserving detailed object contours from the camera stream. Compared to the baseline PointPillars predictions (Figure 1), which occasionally suffered from orientation drift due to sparse point clouds, the refined fusion results leveraged the precise 2D instance masks from YOLOv8 (Figure 2) to better constrain object extents and spatial placement. This complementary interaction allowed the system to maintain consistent detections even in cases of partial occlusion or low LiDAR density (Figure 3), where camera cues provided strong semantic guidance.

Figure 1: Point-Pillars detections
Source: Own

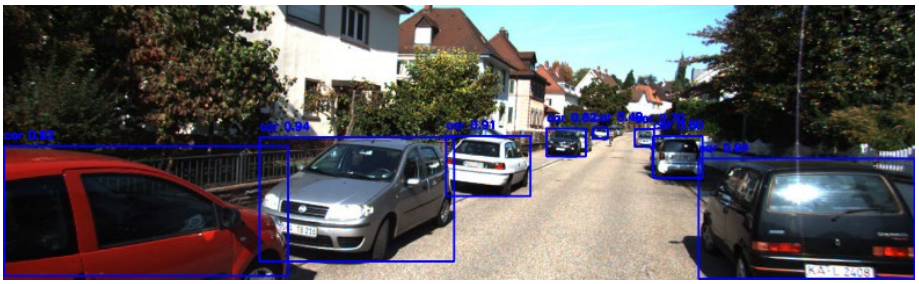

Figure 2: YOLOv8 detections
Source: Own

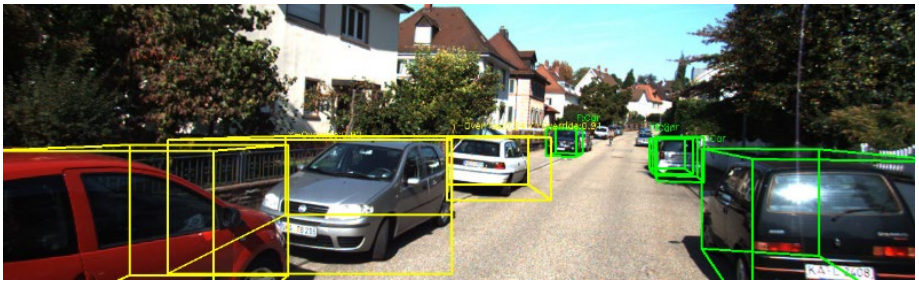

Figure 3: Fusion detections
Source: Own

As a result, the fused detections exhibited improved alignment with true object headings and boundaries, particularly for vehicles parked at non-standard angles and for pedestrians partially occluded by surrounding objects. The synergy between YOLOv8's rich visual segmentation and LiDAR-based geometric clustering proved especially beneficial in cluttered parking scenes, producing more stable and visually coherent 3D bounding boxes. Such improvements are critical for autonomous parking and valet applications, where accurate object orientation and extent directly influence path planning, obstacle avoidance, and overall maneuvering safety.

5 Conclusion and Future Work

This study presented a late-fusion framework combining YOLOv8-seg and PointPillars for 3D object detection, integrating semantic masks with LiDAR geometry to overcome the limitations of single-sensor systems. Evaluation on the KITTI Vision Benchmark Suite demonstrated notable improvements: car precision increased from 59.53% to 70.17%, while the pedestrian F1-score rose from 0.40 to

0.51, highlighting the effectiveness of the rescue phase in maintaining vulnerable road user detections. The use of DBSCAN clustering and PCA-based orientation refinement ensured 3D bounding boxes were spatially accurate and correctly aligned, enhancing reliability for autonomous parking and urban navigation.

A slight drop in recall for cyclists indicates room for improvement, but the overall increase in precision and detection stability supports practical deployment in safety-critical scenarios. Future work will focus on adaptive thresholding to better handle distant or small objects, and on refining the rescue and override mechanisms to maintain robust performance under adverse weather and low-visibility conditions. These enhancements aim to further strengthen the integration of semantic and geometric data, advancing the system toward reliable, real-time perception for autonomous driving and intelligent parking applications.

References

Fei, J., Chen, W., Heidenreich, P., Wirges, S., & Stiller, C. (2020). SemanticVoxels: Sequential Fusion for 3D Pedestrian Detection using LiDAR Point Cloud and Semantic Segmentation. *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 185–190. https://doi.org/10.1109/MFI49285.2020.9235240

He, T., Sun, P., Leng, Z., Liu, C., Anguelov, D., & Tan, M. (2023). *LEF: Late-to-Early Temporal Fusion for LiDAR 3D Object Detection* (arXiv:2309.16870). arXiv. https://doi.org/10.48550/arXiv.2309.16870

Liao, Y., Xie, J., & Geiger, A. (2022). *KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D* (arXiv:2109.13410). arXiv. https://doi.org/10.48550/arXiv.2109.13410

Malić, D., Fruhwirth-Reisinger, C., Prutsch, A., Lin, W., Schulter, S., & Possegger, H. (2025). *GBlobs: Local LiDAR Geometry for Improved Sensor Placement Generalization* (arXiv:2510.18539). arXiv. https://doi.org/10.48550/arXiv.2510.18539

Mitiu, M. A. (n.d.). *3D VISION OBJECT IDENTIFICATION USING YOLOv8*.

Ngo, T. B., Ngo, L., Phi, A. V., Nguyen, T. T. H. T., Nguyen, A., Brown, J., & Perera, A. (2025). C2L3-Fusion: An Integrated 3D Object Detection Method for Autonomous Vehicles. *Sensors*, *25*(9), 2688. https://doi.org/10.3390/s25092688

Pang, S., Morris, D., & Radha, H. (2020). *CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection* (arXiv:2009.00784). arXiv. https://doi.org/10.48550/arXiv.2009.00784

Shlens, J. (2014). *A Tutorial on Principal Component Analysis* (arXiv:1404.1100). arXiv. https://doi.org/10.48550/arXiv.1404.1100

Soyyigit, A., Yao, S., & Yun, H. (2024). VALO: A Versatile Anytime Framework for LiDAR-Based Object Detection Deep Neural Networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *43*(11), 4045–4056. https://doi.org/10.1109/TCAD.2024.3443774

Vora, S., Lang, A. H., Helou, B., & Beijbom, O. (2020). *PointPainting: Sequential Fusion for 3D Object Detection* (arXiv:1911.10150). arXiv. https://doi.org/10.48550/arXiv.1911.10150

Wang, D., Lu, X., & Rinaldo, A. (2019). *DBSCAN: Optimal Rates For Density Based Clustering* (arXiv:1706.03113). arXiv. https://doi.org/10.48550/arXiv.1706.03113

Yao, X., Liu, P., Zhou, J., Wang, Z., Fan, S., & Wang, Y. (2025). MAT-PointPillars: Enhanced PointPillars algorithm based on multi-scale attention mechanisms and transformer. *PLOS ONE*, *20*(6), e0325373. https://doi.org/10.1371/journal.pone.0325373

Zhang, T., Zhang, X., Liao, Z., Xia, X., & Li, Y. (2024). AS-LIO: Spatial Overlap Guided Adaptive Sliding Window LiDAR-Inertial Odometry for Aggressive FOV Variation. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10829–10836. https://doi.org/10.1109/IROS58592.2024.10801561