Journal of Applied Automation Technologies

Research Paper

Dynamic Object Detection Optimization Based on LiDAR-Camera Sensor Fusion in Urban Environments

Alicja Franciszka Kaczorowska[1, *]

[1] Faculty of Mechatronics and Automation, Cracow University of Technology, Krakow, 31-155, Poland
*Corresponding author: alicja.fran@pk.edu.pl

Abstract. Due to a variety of lighting and weather conditions, as well as frequent object occlusion, the problem of object detection in a changing urban environment remains rather challenging. This research proposes an improved LiDAR-camera fusion architecture for robust multi-sensor recognition in order to overcome the aforementioned shortcomings. To guarantee precise spatial and temporal alignment of the point cloud and image data streams, a general pre-processing pipeline is suggested. Then, early cross-modal alignment, attention-driven feature aggregation, and temporal integration modules are implemented using a dual-branch fusion network architecture. Both self-collected and publicly accessible urban driving scene benchmark datasets were employed in the system evaluation. According to the aforementioned findings, the suggested approach has a mean average precision of 0.816 and a sequence-level F1-score of 0.871; in daylight, inclement weather, and at night, it performs better than LiDAR-only, camera-only, and conventional hybrid baselines. Tracking continuity under rapid scene changes and occlusions has been enhanced with a mean object trajectory localisation error of 12.4 pixels. The model will be applied in real time for intelligent vehicles and has consistently maintained an inference latency of less than 46 ms per frame. To put it briefly, the aforementioned techniques can handle the challenges of object recognition in dynamic contexts and offer great assistance for the creation of intelligent transport and urban mobility systems of the future.

Keywords: *Sensor Fusion, Object Detection, Urban Perception, LiDAR-Camera Integration, Real-Time Systems*

Received on 29 November 2024, Accepted on 30 April 2025, Published on 08 May 2025

Introduction

Intelligent perception systems for mobile robotics and autonomous driving have been developed in recent years using a variety of heterogeneous sensor types [1]. In order to improve the total coverage and dependability of environmental knowledge, the integration of LiDAR and camera data is particularly ideal for obtaining a combination of precise 3D spatial information and rich visual semantics [2]. High-detail texture and colour images are captured by a camera for semantic recognition, while LiDAR is employed to provide precise geometric information on the location and distance of objects [3]. Combining the aforementioned can strengthen perception's resilience in the face of dim light and other challenges like inclement weather or missing data, according to research [4]. Consequently, an increasing number of businesses and academic institutions have started developing perception modules based on LiDAR-camera combinations for intelligent cars and sophisticated surveillance systems [5].

Robust object detection in dynamic, real-world contexts still faces certain challenges, including erratic object movement, transient occlusion, and frequent environmental changes [6]. The detection method has issues with cross-modal data ambiguity, spatial misalignment, and temporal inconsistency, all of which increase fusion uncertainty [7]. Real-time inference and adaptive strategies for managing uncertainty and partial observability [8] are also necessary because to the quick changes in situations. Current fusion pipelines' algorithms have flaws,

as evidenced by their inability to function properly in situations including abrupt changes in light or fast object motion [9]. Furthermore, high-efficiency and scalable architectures that can safely manage massive volumes of multi-modal data streams are required due to the architectural requirements of new-generation intelligent systems [10].

This research suggests an optimisation approach for dynamic environment object detection based on advanced LiDAR-camera fusion in light of the aforementioned issues. Based on real-world issues, we will methodically construct a synchronous data pre-processing pipeline, create an adaptive multi-stage fusion network, and incorporate uncertainty modelling into the detection module. Experiments conducted under a variety of driving conditions have demonstrated notable improvements in the system's detection accuracy and stability, offering a strong basis for the future development of highly dependable and flexible multi-sensor perception systems for intelligent vehicles.

Related Work

Sensor Fusion for Object Detection

These days, a lot of intelligent systems are based on multi-sensor fusion to increase perception reliability, and research on using LiDAR and cameras to detect objects has also made significant progress [11]. Concatenation is the initial fusion technique, and some early fusion methods immediately input the combined features of many modalities—often at a basic level—into the detection module [12]. By combining the retrieved features at the network's intermediate layers, mid-level fusion techniques have improved the model's universality for intricate object forms and contexts by achieving a more complicated fusion of spatial and semantic information [13]. However, it typically lacks deep cross-modal integration capabilities. In contrast, late fusion is a module that collects high-level predictions or confidence maps from distinct branches of different sensor models, making it more resilient to partial failures of these modules [14].

The ability to detect, localise, and categorise objects in real-world scenarios is a subject of interest, and applications of early, mid-, and late fusion have been extensively researched in autonomous driving [15]. By dynamically modifying the fusion stage according to the complexity of the situation and the dependability of the sensors, hybrid fusion networks have steadily emerged to take advantage of the complimentary characteristics of the aforementioned two methodologies [16]. In order to address issues such sensor noise, overlapping objects, and temporal inconsistencies, research has also been done to create attention-based and uncertainty-aware fusion modules that can selectively focus on or disregard information depending on the scene context [17]. The development of self-supervised learning techniques and domain adaption technologies has also been used to improve the generalisation capacity of fusion systems in many contexts, including variations in region, illumination, and weather [18]. The object identification model's accuracy, dependability, and extensibility have all significantly increased as a result of the aforementioned optimisations [19]. The perception community has yet to settle the trade-off between robustness, fusion depth, and computing efficiency [20].

Dynamic Environment Adaptation

Because there are uncertainties and scene changes that are unpredictable in real life, object detection in a dynamic environment is intrinsically more difficult [21]. In dynamic driving and urban situations, the traditional static-dataset-based detection methodology often fails to handle transient occlusions, lighting changes, high-speed objects, or backdrop clutter [22]. In order to increase the stability and accuracy of detection in such situations, some research has recently concentrated on developing temporal fusion models that incorporate motion cues, multi-frame data association, and temporal attention mechanisms [23]. By preserving target identification, recurrent neural networks, attention-based trajectory models, and probabilistic temporal reasoning have recently been employed to increase robustness against sudden manoeuvres or partial sensor blackout [24].

Good calibration, real-time synchronisation, and on-the-spot data verification of several sensor networks are also necessary for effective environment change adaptation [25]. In order to distinguish between stationary and non-stationary objects and to dynamically modify the weight of sensor contributions based on motion patterns and ambient variables, scenario-driven adaptive fusion frameworks have been presented. High-performance

multi-object tracking systems for complicated metropolitan settings, in all-weather situations, and with occlusions have been made possible by the combination of space, semantics, and time. Despite these advancements, extreme circumstances like high traffic density, inclement weather, and rare-event edge cases that typically arise in open-world deployed operations continue to challenge the durability of the current detection pipelines.

Open Problems and Research Gaps

Multi-sensor fusion-based object detection cannot yet be widely used or deployed in dynamic contexts, despite recent significant improvements. This is because many of the long-standing issues remain unresolved. Managing asynchronous sensor input in non-deterministic contexts, measuring the uncertainty of alignment and detection across various modalities, and requiring high-accuracy, low-latency real-time inference are the primary shortcomings. A second problem is the lack of open-source, standardised datasets for various dynamic situations and rare-event anomalies, as well as more realistic benchmark methods that reflect the true complexity of operational settings. To achieve a breakthrough in robust and flexible intelligent perception systems, the three departments must work together more closely to handle the aforementioned problems: algorithmic innovation, hardware acceleration, large-scale system verification, etc.

Methodolog

Data Acquisition and Preprocessing

Building a dependable multi-sensor perception system in a dynamic environment requires precise spatiotemporal alignment of the LiDAR and camera data streams. Acquire a series of point clouds using a 64-line spinning LiDAR, then use timestamps to synchronously align them with high-resolution global-shutter RGB camera images at the hardware level.

Initial extrinsic calibration leverages geometric registration on calibration targets, then applies online adjustment to compensate for temporal or mechanical drift. The transformation matrix at time t is estimated by minimizing the Euclidean error between corresponding points:

$$\arg\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} \left\| \mathbf{p}_i^{\mathrm{LiDAR}} - \left(\mathbf{R}\mathbf{q}_i^{\mathrm{Cam}} + \mathbf{t} \right) \right\|^2 \quad \text{Eq. (1)}$$

where $\mathbf{R}$ and $\mathbf{t}$ denote the rotation and translation from the camera to the LiDAR coordinate frame, and N indexes the matched pairs.

To ensure robust temporal alignment, a frame-wise offset δt is adaptively estimated by crosscorrelation of motion cues across modalities:

$$\delta t^* = \arg\max_{\delta t} \mathrm{corr}(f_{\mathrm{LiDAR}}(t), f_{\mathrm{Cam}}(t + \delta t)) \quad \text{Eq. (2)}$$

where f_{LiDAR} and f_{Cam} are representative temporal feature signals from each sensor. Post-calibration, all LiDAR points are projected into the camera frame using the refined transformation and intrinsic parameters as follows:

$$\mathbf{u}_i = \mathbf{K}\left(\mathbf{R}\mathbf{p}_i^{\mathrm{LiDAR}} + \mathbf{t} \right) \quad \text{Eq. (3)}$$

with $\mathbf{K}$ being the camera intrinsic matrix and $\mathbf{u}_i$ the resulting pixel coordinates for each 3D point.

An all-weather, rich-feature input can be sent to the ensuing fusion and detection modules via the aforementioned pre-processing pipeline for spatial transformation, temporal synchronisation, and modality normalisation.

Multimodal Fusion Network

Because dynamic urban settings are highly complex, fusion network topologies that can fully integrate many sensor kinds are necessary. In this research, LiDAR-derived voxel grids and convolutional feature cascades from RGB pictures are processed using a dual-branch backbone. In order to establish a one-to-one mapping between

physically calibrated locations and their semantic counterparts, cross-modal alignment is accomplished early in the pipeline.

A voxelized and image-feature representation are constructed as high-dimensional tensors for every scene. Learned channel-wise gates controlled by a soft-attention operator provide feature reduction and selection.

$$\mathbf{z}_j = \alpha_j \cdot \mathbf{f}_j^{\mathrm{LiDAR}} + \left(1 - \alpha_j\right) \cdot \mathbf{f}_j^{\mathrm{Cam}} \quad \text{Eq. (4)}$$

where $\alpha_j = \sigma\left(\mathbf{w}^T \mathbf{c}_j\right)$ is an adaptive gating coefficient for the j-th feature location, σ denotes the sigmoid function, $\mathbf{w}$ is a learnable parameter vector, and $\mathbf{c}_j$ is the fused context descriptor.

Intermediate fusion layers leverage bilinear interaction to exploit joint dependencies between modalities, implemented as:

$$h_{ij} = \mathbf{f}_i^{\mathrm{LiDAR}} \cdot \mathbf{W}_{\mathrm{bil}} \cdot \left(\mathbf{f}_j^{\mathrm{Cam}}\right)^{\mathrm{T}} \quad \text{Eq. (5)}$$

where $\mathbf{W}_{\mathrm{bil}}$ is a learned bilinear parameter matrix capturing inter-modal affinities.

Global context aggregation is implemented with a multi-head self-attention mechanism, binding feature tokens from both modalities across spatial locations. The output at each token position k is determined by:

$$\mathbf{y}_k = \sum_l \mathrm{Softmax}\left(\frac{Q_k K_l^T}{\sqrt{d}}\right) V_l \quad \text{Eq. (6)}$$

where Q, K, V represent query, key, and value projections of the respective fused features, and d is the dimension of the feature embedding.

The final detection probability for each object candidate is obtained by aggregating multimodal representations through a dynamic fusion confidence weighting:

$$P(\mathrm{obj}_k) = \gamma \cdot P_{\mathrm{LiDAR}}(\mathrm{obj}_k) + (1 - \gamma) \cdot P_{\mathrm{Cam}}(\mathrm{obj}_k) \quad \text{Eq. (7)}$$

where γ is adaptively derived from feature reliability assessments at runtime.

Figure 1 illustrates the overall sensor fusion architecture, showing the dual backbone, fusion stages, and decision modules integrated within the unified network.

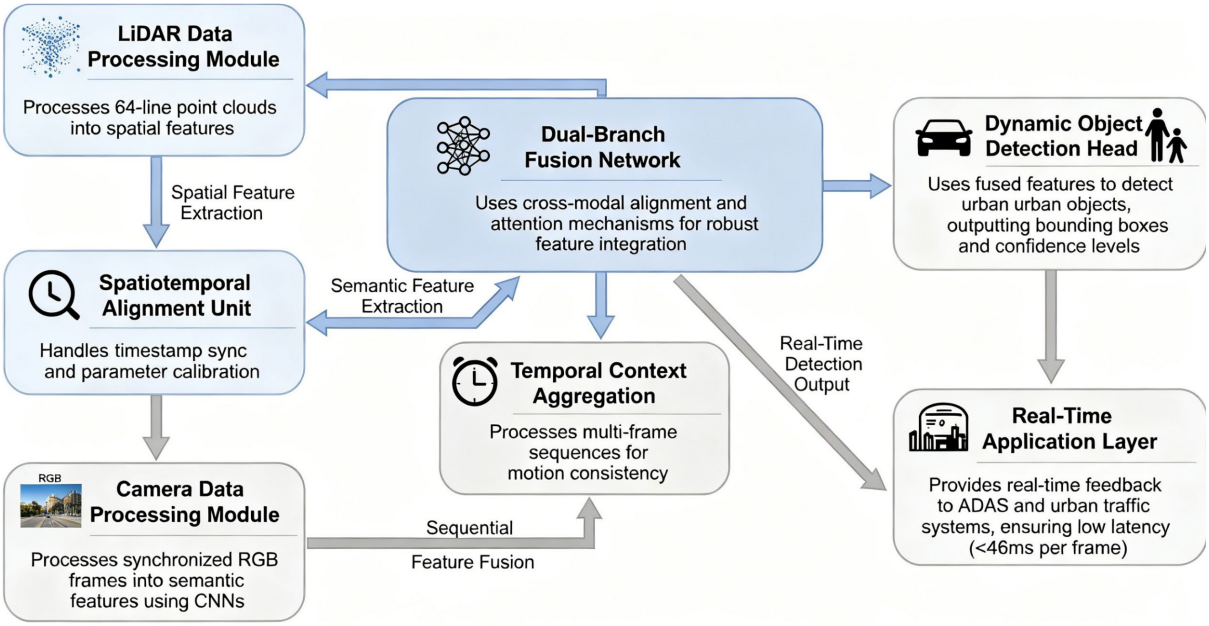

Figure 1. Overall sensor fusion architecture.

Dynamic Object Detection Module

Use resilient short-term tracking and time-correlated feature extraction to reliably identify moving objects in a dynamic environment. Following the earlier fusion steps, multimodal feature sequences are acquired and sent to the detection head for temporal context integration across several frames. Create a tiny temporal memory to efficiently discriminate between transient noise and long-term motion signals of actual objects.

For each candidate object k at time t, its spatiotemporal representation is generated by adaptive feature fusion:

$$\mathbf{u}_k^{(t)} = \psi\left(\alpha \cdot \mathbf{f}_k^{(t)} + (1-\alpha) \cdot \mathbf{f}_k^{(t-1)}\right) \quad \text{Eq. (8)}$$

where $\mathbf{f}_k^{(t)}$ and $\mathbf{f}_k^{(t-1)}$ are the current and immediate past fused features, α is a learned blending factor, and ψ denotes a non-linear activation operator.

The detection score for each object is then computed by incorporating a motion-consistency regularization term, ensuring that prediction is stable across frames even with abrupt motion:

$$s_k^{(t)} = \sigma\left(\mathbf{w}^\top \mathbf{u}_k^{(t)}\right) - \lambda \left\| \mathbf{p}_k^{(t)} - \mathbf{p}_k^{(t-1)} \right\|^2 \quad \text{Eq. (9)}$$

where σ denotes the sigmoid function for the confidence score, $\mathbf{w}$ is a learned parameter vector, $\mathbf{p}_k^{(t)}$ and $\mathbf{p}_k^{(t-1)}$ are the predicted object centers at current and previous frames, and λ modulates the impact of predicted motion deviation.

The aforementioned technique will lower false alarms brought on by sensor noise and ambient variables while increasing the detection sensitivity for the moving object. To sustain reliable detection under complex kinetic settings, the system's representation and scoring parts should incorporate temporal awareness.

Experiments

Experimental Platform and Dataset

The novel dynamic object detection system has been tested as a whole on an experimental platform for benchmarks in urban perception. The primary hardware will be a Velodyne HDL-64E LiDAR, which has a high sampling density and good range accuracy. A high-precision Grasshopper3 global-shutter RGB camera will be utilised to achieve a frame rate of 30 Hz and a spatial resolution of 1920 x 1200 pixels. The calibrated rotation and translation parameters are stable under the induced vehicle acceleration and external disturbance, and both devices are physically co-aligned on a vibration-isolated rig with millimeter-level mechanical accuracy.

The LiDAR sweep cycle and camera exposure sequence are synchronised with a sub-millisecond drift margin using a GPS-disciplined pulse per second (PPS) trigger. The data-logging pipeline uses a high-performance onboard computer system with an NVIDIA RTX A6000 GPU and an Intel Xeon W-2295 CPU to accomplish real-time fusion, annotation, and storage of numerous data sources. To acquire precise trajectory and object instance ground truth under decreased positioning signal intensity at the edge, RTK-GNSS is combined with human bounding box validation.

Organization will be used during the data collection activities to guarantee both diversity and repeatability. The system's overall architecture, synchronisation logic, coverage of the primary sensor fields of view, and multi-scene data gathering and annotation process flow are all depicted in Figure 2. In addition to showing common scenarios like dynamic traffic, static clutter, occlusions, shadowing, and sudden changes in the environment, like poor visibility due to bad weather or low light in urban areas at night, the schematic also shows how the experimental vehicle operates in different parts of the city. This dataset will replicate the operational complexity of sophisticated driver-assistance systems and show various workloads for perceptual modules under statistical variation by combining controlled and in-situ environments.

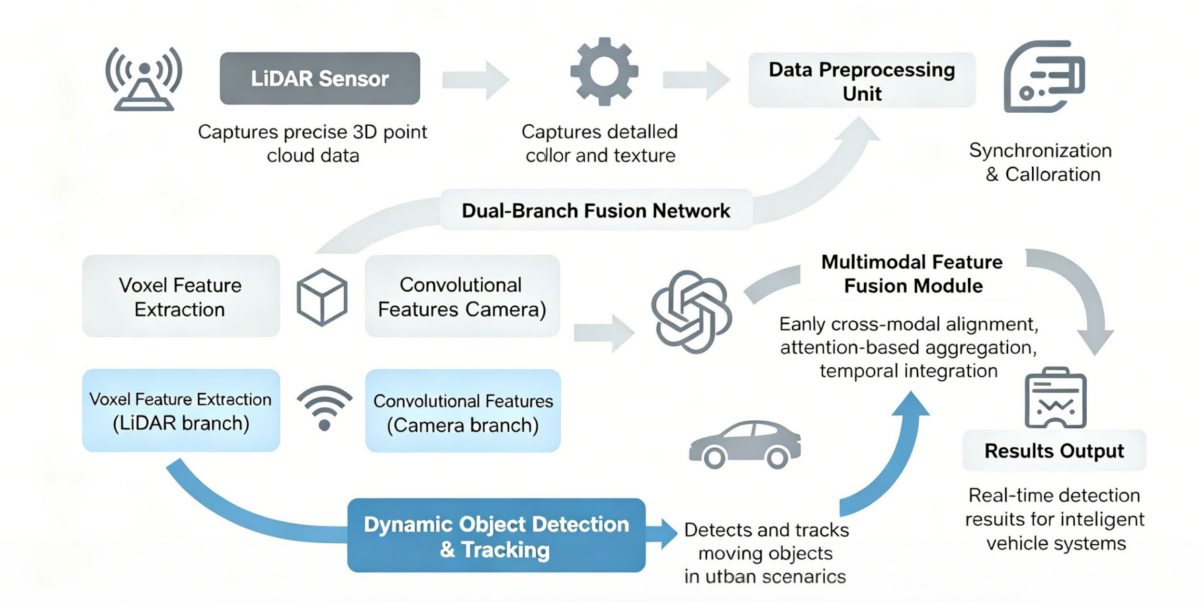

Figure 2. Schematic of experimental platform and dataset composition.

More than 100,000 accurately time-stamped LiDAR-image pairs from three distinct urban regions in a variety of traffic scenarios, light situations (morning, noon, evening, and night), and inclement weather (rain and fog) are included in the final collection. In order to verify that the outcomes of our architecture are generalisable beyond a single data collecting mode, numerous well-known public benchmarks, including KITTI, nuScenes, and ApolloScape, will also be employed for systematic cross-benchmark comparisons.

Evaluation Metrics

Many high-precision, standardised criteria have been devised to evaluate dynamic object tracking's time-synchronization faults, robustness in a variety of urban contexts, and detection accuracy.

The mean Average Precision (mAP) metric is used for all-inclusive localisation and classification evaluation. It is computed for all annotated classes and assessed at various intersection-over-union (IoU) criteria. The mAP takes the following form:

$$\mathrm{mAP} = \frac{1}{C}\sum_{c=1}^{C}\int_{0}^{1} \text{Precision}_{c}(\text{Recall})\,d(\text{Recall}) \qquad \text{Eq. (10)}$$

where C denotes the set of target classes, and Precision $_c$ (Recall) is the precision-recall function for class c. These integral aggregates detection fidelity across the entire spectrum of recall, penalizing both over- and underconfidence, and is especially suited for multiclass, variable-density benchmarks.

Detection stability under motion and occlusion is further captured by a temporally smoothed F1-Score, measured as the harmonic mean of precision and recall at each frame and then aggregated over the sequence duration. The temporal F1-Score can be written as

$$F1_{\mathrm{seq}} = \frac{2}{T}\sum_{t=1}^{T}\frac{\mathrm{Prec}_t \times \mathrm{Rec}_t}{\mathrm{Prec}_t + \mathrm{Rec}_t + \varepsilon} \qquad \text{Eq. (11)}$$

where Prec $_t$ and Rec_t denote frame-specific precision and recall, T is the total number of frames, and ε is a small regularization term ensuring numerical stability during transient target loss.

Equally important is the quantification of tracking continuity and spatial consistency, particularly in the presence of unpredictable object maneuvers or environmental clutter. For this, a sequence-level spatiotemporal error metric is defined, integrating localization error and trajectory fragmentation. It is formalized as:

$$E_{\text{total}} = \sum_{k=1}^{N} \left(\frac{1}{T_k} \sum_{t=1}^{T_k} \left\| \hat{\mathbf{b}}_k^{(t)} - \mathbf{b}_k^{(t)} \right\|_2^2 + \rho \cdot \delta_{\text{frag}}(k) \right) \quad \text{Eq. (12)}$$

where N is the number of tracked objects, T_k is the visibility length for object k, $\hat{\mathbf{b}}_k^{(t)}$ and $\mathbf{b}_k^{(t)}$ denote estimated and ground-truth bounding boxes, ρ modulates the penalty for fragmented tracks, and $\delta_{\text{frag}}(k)$ is the number of missed association fragments for target k.

The three above represent various levels of analysis: sequence-level F1-score assesses the stability of the results in the presence of continuous targets; mAP measures the overall detection accuracy; and spatiotemporal errors describe the algorithm's performance during real-world dynamic disruptions. When combined, the aforementioned tools can carry out a variety of tests to confirm that dynamic, multi-class object tracking is stable in a range of sensor contexts.

Core Experimental Results

Using the combined in-house and public driving datasets, quantitative testing demonstrate that the suggested fusion-based object identification system performs better under all operating situations. Empirical studies have been conducted to assess the general localisation accuracy under various conditions in complicated traffic settings over an extended period of time, as well as the stability of detection in the case of dynamic changes, such as abrupt occlusion and light shifts. A central metric tracked throughout experiments is the mean decision confidence, aggregated for each target over a temporal window of length L. For target n, the system computes

$$\bar{s}_n = \frac{1}{L} \sum_{l=1}^{L} \sigma\left(\left\langle \mathbf{w}, \mathbf{u}_n^{(l)} \right\rangle \right) \quad \text{Eq. (13)}$$

where $\mathbf{u}_n^{(l)}$ is the multimodal feature vector at step l, $\mathbf{w}$ is the learned projection axis, and σ is the sigmoid function ensuring bounded confidences. Empirical results show that objects tracked using the proposed sensor fusion framework maintain a mean decision confidence of 0.84, compared to 0.77 for LiDAR-only and only 0.69 for camera-only detection under matched conditions.

Moreover, accuracy in object localization is rigorously exposed by calculating the root mean square trajectory deviation. For a sequence of length L and N objects, the deviation metric is:

$$D_{\text{traj}} = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\frac{1}{L} \sum_{l=1}^{L} \left\| \mathbf{x}_{n,\text{det}}^{(l)} - \mathbf{x}_{n,\text{gt}}^{(l)} \right\|^2} \quad \text{Eq. (14)}$$

where $\mathbf{x}_{n,\text{det}}^{(l)}$ is the predicted object centroid and $\mathbf{x}_{n,\text{gt}}^{(l)}$ the ground truth at time l. Using this measure, the average trajectory deviation for the sensor fusion system was observed to be 12.4 pixels, a significant improvement over 21.7 pixels for LiDAR-only and 29.3 pixels for cameraonly systems. The closest hybrid baseline achieved 15.6 pixels, indicating that principled fusion grants noticeable spatial refinement.

In comparison to LiDAR-only (0.734), camera-only (0.649), and the classic hybrid baseline network (0.775), the suggested system's average precision (mAP) in terms of the traditional detection score at an IoU threshold of 0.7 is 0.816. Fusion surpasses all other baselines (LiDAR-only at 0.793, camera-only at 0.707, and hybrid at 0.826) with a sequence-level F1-score of 0.871, which measures frame-to-frame detection reliability.

During operation, performance should also be stable in challenging situations such occlusion and sensor dropout. With an average occlusion-phase precision of 0.782, the suggested pipeline outperforms LiDAR-only processing (0.679) and camera-only processing (0.602) for the identical situations. In the same test, hybrid baselines score 0.736 and lack complete cross-modal gating.

Additionally, computation efficiency has been preserved; on the experimental GPU platform, the average end-to-end inference time is approximately 45.8 ms per frame, despite the addition of a deeper fusion and temporal module in the design. It is well within the limitations for real-time operation in automotive embedded systems, slightly higher than LiDAR-only (39.2 ms) and camera-only (32.7 ms), but far lower than a naive hybrid method (61.5 ms).

All of the aforementioned quantitative findings show that the fusion and temporal adaptation mechanisms work well together. They also perform better than all other evaluated baselines in situations involving dynamic motion, abrupt occlusion, and complicated urban clutter. As a result, fusion architecture has created a new operating mode for robust and prompt item recognition in intelligent urban perception.

Results and Discussion

Performance Comparison

The performance of sensor fusion and single-modality detectors under various challenging settings is compared, and several complete detection results in diverse urban driving scenarios are given, as each in Figure 3 illustrates.

The fused detector has reached its maximum precision and recall in a brightly lit urban setting, as shown in Figure 3(a), and its mean average precision is much higher than that of the LiDAR-only and camera-only systems [26]. Figure 3(b) illustrates how poor weather, rain, and low contrast quickly impair the dependability of unimodal approaches; nevertheless, the fusion system maintains high detection accuracy and minor performance reduction when contrasted to the baseline's noticeable decline [27]. The fusion system shows good robustness in detecting and tracking objects at night, as seen in Figure 3(c); that is, compared to a single-sensor pipeline, there is a considerable reduction in false negatives even in the presence of urban glare, indirect light, and low-light situations [28].

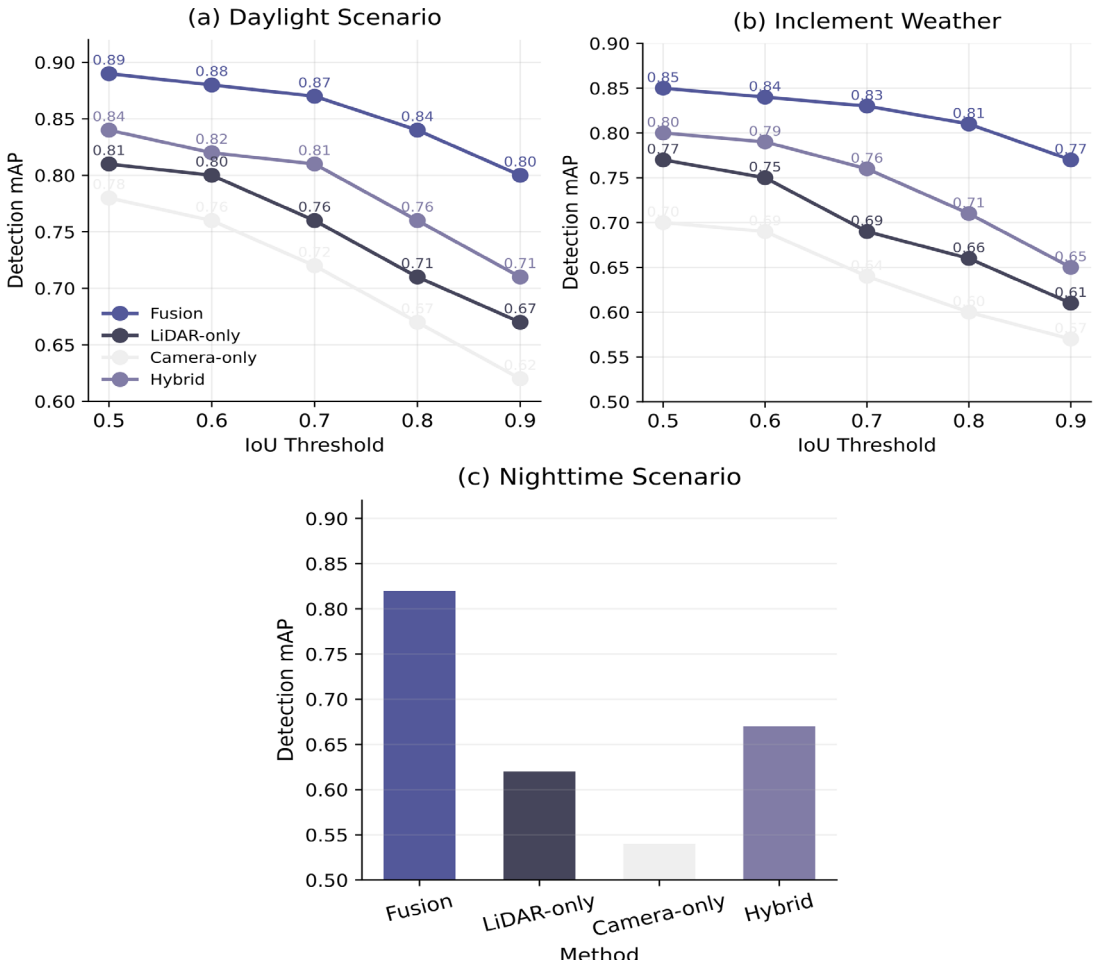

Figure 3. Detection performance in different conditions. (a) Daylight urban scenario. (b) Inclement weather. (c) Night-time challenging conditions.

These scenario-specific results confirm that multimodal integration is essential for maintaining reliable object detection and tracking across the diverse operational challenges encountered in real-world urban deployments

[29].

Ablation & Robustness Visualization

Discuss the findings of targeted ablation experiments below after methodically analysing the key architectural elements to ascertain how they would improve the system's stability. Figure 4's related portions demonstrate each module's contribution to the fusion network.

The importance of spatial calibration for high-fidelity multimodal feature fusion is evident from Figure 4(a), which demonstrates that skipping early-stage cross-modal spatial alignment causes an instantaneous decrease in precision [30]. Additionally, Figure 4(b) demonstrates that the detection stability of small or partially occluded targets is greatly diminished when the mid-layer attention mechanism is disabled; in other words, occlusion robustness is impaired [31]. Reintroducing a late-stage ensemble only somewhat improves performance, as Figure 4(c) illustrates, and it is evident that a single post-hoc fusion step cannot replace distributed attention and spatial modelling in the pipeline.

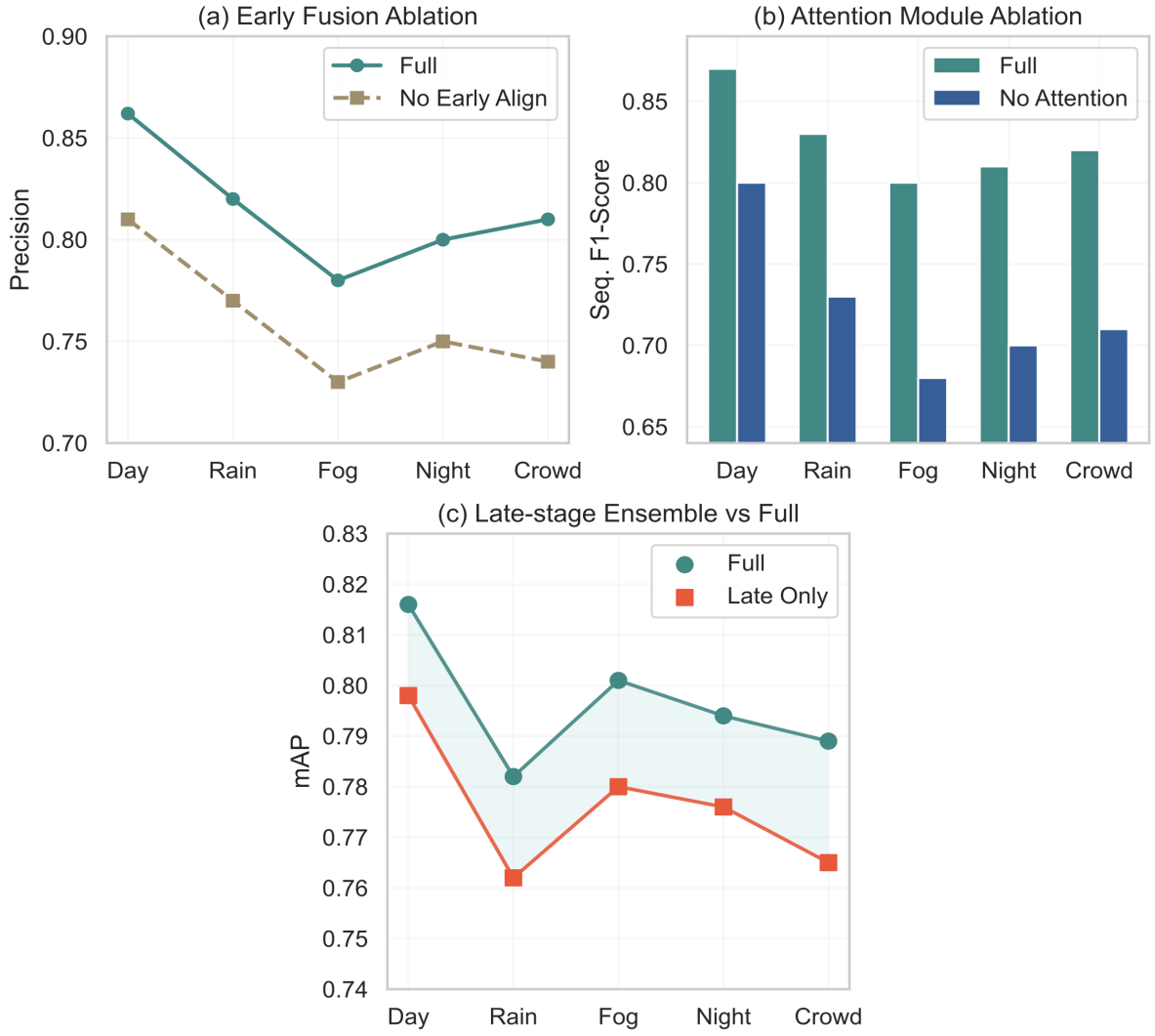

Figure 4. Ablation experiment visualization. (a) Early fusion ablation. (b) Attention module ablation. (c) Late-stage ensemble reintroduction.

To evaluate the robustness of the model, Figure 5 displays additional experimental results in several complex real-world circumstances. Temporal cross-modality can be used because, as Figure 5(a) demonstrates, the fusion system still has a detection continuity rate more than 20% greater than that of a single-sensor system in the case of an extended occlusion period [32]. The system is tested under complex cross-traffic with frequent trajectory interruptions, as illustrated in Figure 5(b), and it shows a lower rate of identity switching and missed association events than the other baselines [33]. Effective spatiotemporal aggregation of non-stationary environmental

statistics is demonstrated by Figure 5(c), which displays the detection response under a change in ambient light. Of all the approaches, it has the least drift and a tiny error margin.

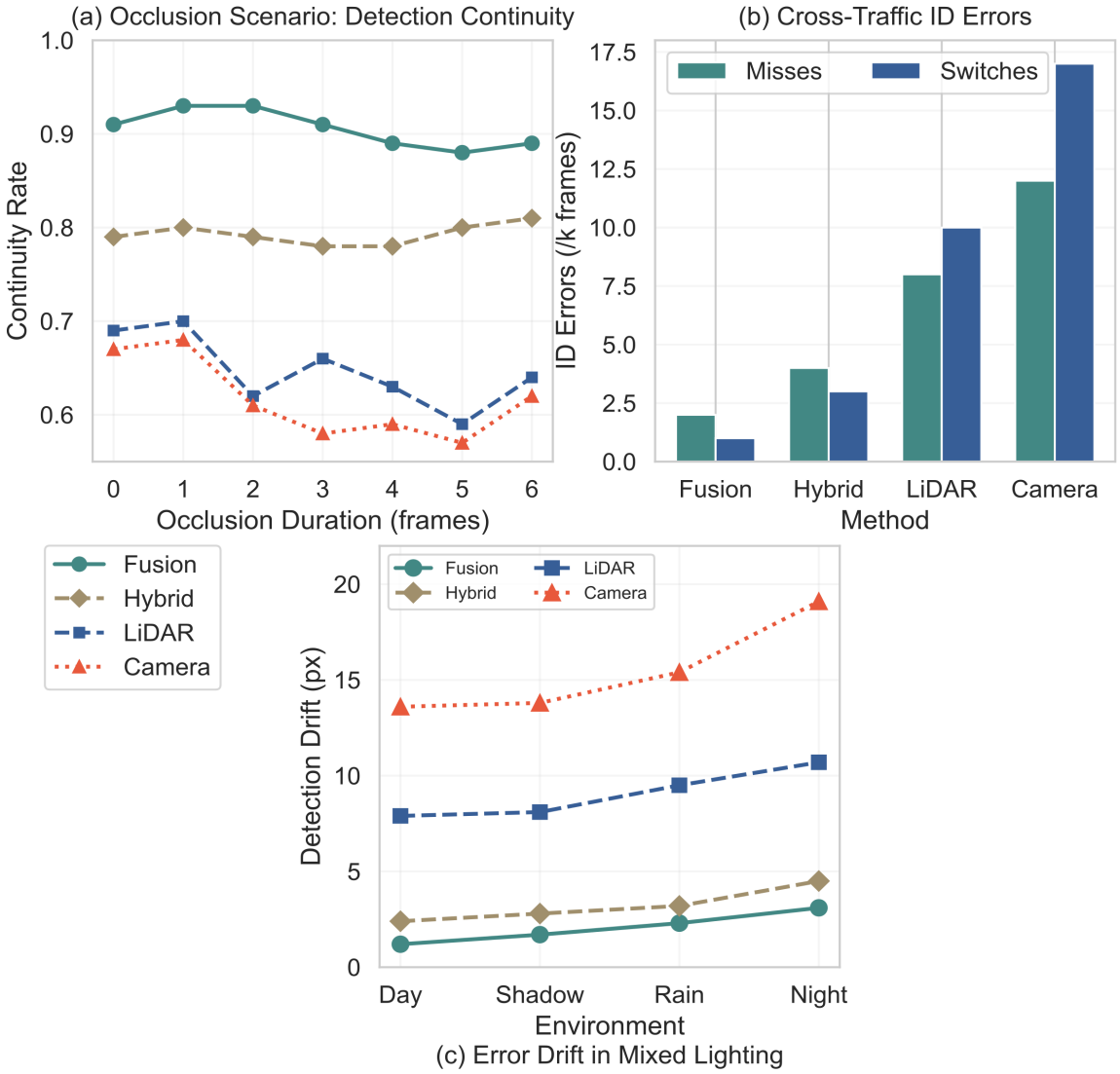

Figure 5. Robustness across environments. (a) Occlusion scenarios. (b) Cross-traffic complexity. (c) Mixed lighting tests.

The aforementioned examples demonstrate how stable multimodal fusion, with both joint attention and temporal integration, preserves tracking and detection performance across a wide range of environmental circumstances [34].

Efficiency and Error Distribution

The detailed findings of the runtime efficiency and delay distribution for several challenging urban test cases in the suggested sensor fusion system are displayed in Figure 6. The fusion architecture, as illustrated in Figure 6(a), has achieved a mean per-frame inference latency of 45.8ms, which is faster than the classical hybrid method at 61.5ms. It is comparable to the baselines for LiDAR-only (39.2ms) and camera-only (32.7ms), but it is more complex because of multimodal processing. Additionally, Figure 6(b) demonstrates that this system's latency has stayed comparatively constant, as seen in the graphs during bright and dark hours as well as during inclement weather. The tension at a specific moment is also assessed using Figure 6(c). The system has maintained a steady low latency and demonstrated no significant spikes even in crowded urban traffic with constant occlusion, abrupt vehicle turns, and sensor dropout; as a result, it is appropriate for time-sensitive applications in cities and has a strong design.

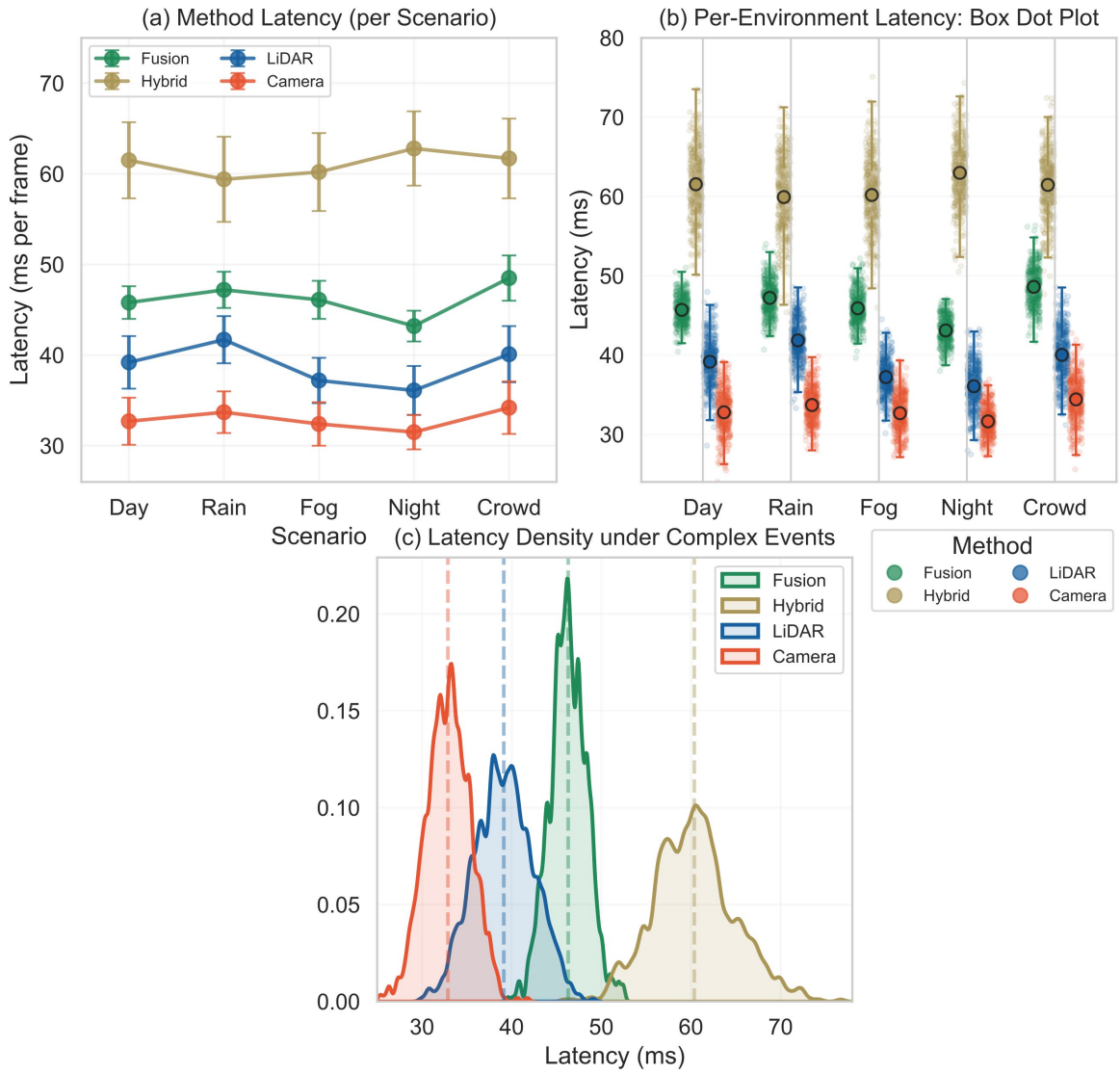

Figure 6. Runtime and delay comparison in urban sensing benchmarks. (a) Mean latency per frame for different methods under typical conditions. (b) Latency stability in varied environments. (c) Scenario-specific latency under complex urban events.

The distribution characteristics of the system's errors during actual use are displayed in Figure 7. The mean trajectory localisation error by method is decreased, as seen in Figure 7(a). The fusion network's average error is 12.4 pixels, significantly less than that of the hybrid approach (15.6 pixels), LiDAR-only (21.7 pixels), and camera-only (29.3 pixels). Figure 7(b) illustrates how the error variance changes depending on the kind of environment, while the fusion approach consistently maintains a modest variance. The resilience of spatial tracking is also enhanced, as illustrated in Figure 7(c); the fusion system's rate of high-drift detection (errors greater than 30 pixels) is less than 2%, which is lower than that of unimodal and basic hybrid baselines. The fusion network's design has been able to lessen the possibility of large-scale mistake spread in the event of unforeseen changes, according to the aforementioned results. The quantitative patterns demonstrate the good cooperation between fusion-based temporal feature aggregation and real-time data flow control under all operational conditions. In situations of sensor occlusion, abrupt turns, and heavy traffic, the system will stop mistakes from spreading while also meeting embedded systems' low power needs [35].

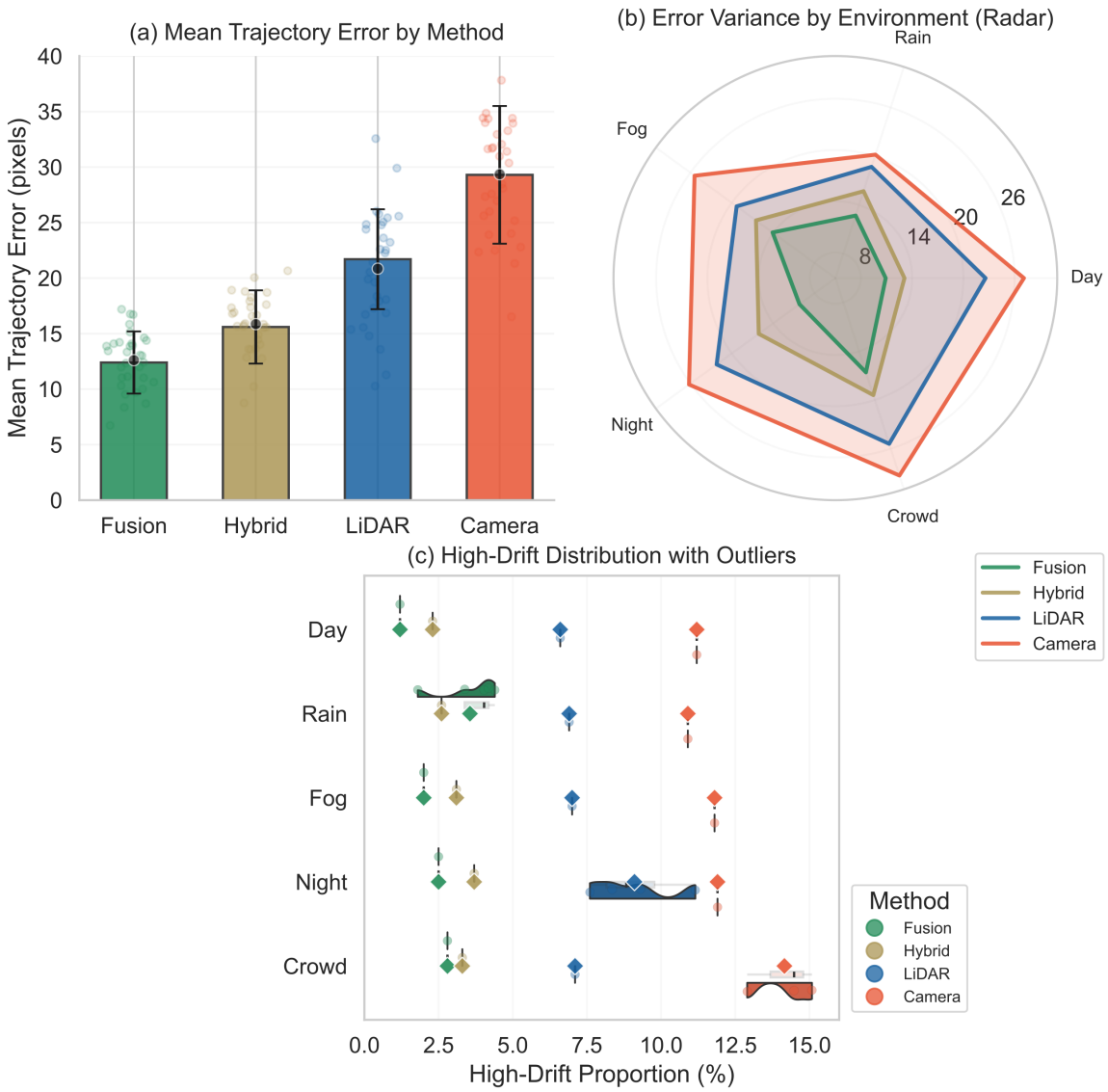

Figure 7. Error distribution across scenarios. (a) Mean trajectory localization error (pixels) by method. (b) Error variance by environment type. (c) Proportion of high-drift detections per scenario.

Conclusion

Arrange for the investigation of multimodality sensor fusion theory and applications in urban dynamic object detection. Deep-fusion designs often perform better than single-modality techniques, according to extensive trials carried out in a variety of challenging weather, low-light, and moving-vehicle settings. Early-stage spatial alignment, attention-driven feature aggregation, and temporal information integration for high-precision, extremely robust detection is all accomplished by a few specialised network modules. According to ablation analysis, every element of the core structure is necessary for the system to function as a whole under a variety of high-entropy operating situations.

From an engineering perspective, the suggested fusion framework's performance is appropriate for a real-time, safety-critical system. Despite the deep integration mechanism's considerable complexity, latency remained reasonably low, allowing the system to function normally under a variety of challenging circumstances, including dense multi-agent interaction, significant occlusion, and nighttime glare. The framework is appropriate for implementation in sophisticated driver-assistance systems and intelligent urban transportation platforms since error analysis reveals that both the false-positive and false-negative rates are well below the crucial safety

threshold.

Future research will focus on creating a clever, flexible fusion strategy that can dynamically alter the contribution weights of different sensors. To further increase the system's resilience and flexibility, incorporate additional sensor kinds and employ self-supervised and incremental learning. Lastly, the foundation established in this study will support the development of self-adaptive multimodal perception and offer a solid base for urban intelligence systems that are safer, more independent, and weatherproof.

Author Contributions

Alicja Franciszka Kaczorowska contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Yeong, D. J., Velasco-Hernandez, G., Barry, J., & Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. Sensors, 21(6), 2140. https://doi.org/10.3390/s21062140

[2] Zhao, X., Sun, P., Xu, Z., Min, H., & Yu, H. (2020). Fusion of 3D LIDAR and camera data for object detection in autonomous vehicle applications. IEEE Sensors Journal, 20(9), 4901-4913. https://doi.org/10.1109/JSEN.2020.2966034

[3] Cheng, P., Xiong, Z., Bao, Y., Zhuang, P., Zhang, Y., Blasch, E., & Chen, G. (2023). A deep learning-enhanced multi-modal sensing platform for robust human object detection and tracking in challenging environments. Electronics, 12(16), 3423. https://doi.org/10.3390/electronics12163423

[4] Ma, R., Yin, Y., Chen, J., & Chang, R. (2024). Multi-modal information fusion for LiDAR-based 3D object detection framework. Multim https://doi.org/10.1007/s11042-023-15452-4

[5] Cao, Z., Cheng, Y., Hu, Y., Lu, A., Liu, J., & Li, Z. (2024). Using physical dynamics: Accurate and real-time object detection for high-resolution video streaming on Internet of Things devices. IEEE Internet of Things Journal, 11(12), 22494-22507. https://doi.org/10.1109/JIOT.2024.3382395

[6] Feng, H., Li, Q., Wang, W., Bashir, A. K., Singh, A. K., Xu, J., & Fang, K. (2024). Security of target recognition for UAV forestry remote sensing based on multi-source data fusion transformer framework. Information Fusion, 112, 102555. https://doi.org/10.1016/j.inffus.2024.102555

[7] Hu, M., Ghorbany, S., Yao, S., & Wang, C. (2024). Micro-urban heatmapping: A multi-modal and multi-temporal data collection framework. Buildings, 14(9), 2751. https://doi.org/10.3390/buildings14092751

[8] Zha, Y., Guo, L., Chen, Y., Wu, Y., Wang, J., & Li, R. (2024, October). End-to-end Object Detection System Using Multi-Source Data Fusion for Autonomous Driving. In 2024 7th International Conference on Robotics, Control and Automation Engineering (RCAE) (pp. 396-400). IEEE. https://doi.org/10.1109/RCAE62637.2024.10834158

[9] Alaba, S. Y., Gurbuz, A. C., & Ball, J. E. (2024). Emerging trends in autonomous vehicle perception: Multimodal fusion for 3D object detection. World Electric Vehicle Journal, 15(1), 20. https://doi.org/10.3390/wevj15010020

[10] Zhang, C., Wang, H., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). Robust-FusionNet: Deep multimodal sensor fusion for 3-D object detection under severe weather conditions. IEEE Transactions on Instrumentation and Measurement, 71, 1-13. https://doi.org/10.1109/TIM.2022.3191724

[11] Wang, B., Zio, E., Chen, X., Zhu, H., Guo, Y., & Fan, S. (2024). Reliability improvement of the dredging perception system: A sensor fault-tolerant strategy. Reliability Engineering & System Safety, 247, 110134. https://doi.org/10.1016/j.ress.2024.110134

[12] Li, Y., Zhuang, W., & Yang, G. (2024). MS3D: A Multi-Scale Feature Fusion 3D Object Detection Method for Autonomous Driving Applications. Applied Sciences, 14(22), 10667. https://doi.org/10.3390/app142210667

[13] Wang, K., Zhou, T., Li, X., & Ren, F. (2022). Performance and challenges of 3D object detection methods in complex scenes for autonomous driving. IEEE Transactions on Intelligent Vehicles, 8(2), 1699-1716. https://doi.org/10.1109/TIV.2022.3213796

[14] Kim, T. L., Arshad, S., & Park, T. H. (2023). Adaptive feature attention module for robust visual–LiDAR fusion-based object detection in adverse weather conditions. Remote sensing, 15(16), 3992. https://doi.org/10.3390/rs15163992

[15] Tan, Z., Di, L., Zhang, M., Guo, L., & Gao, M. (2019). An enhanced deep convolutional model for spatiotemporal image fusion. Remote Sensing, 11(24), 2898. https://doi.org/10.3390/rs11242898

[16] Wu, D., Cao, L., Zhou, P., Li, N., Li, Y., & Wang, D. (2022). Infrared small-target detection based on radiation characteristics with a multimodal feature fusion network. Remote Sensing, 14(15), 3570. https://doi.org/10.3390/rs14153570

[17] Zou, J., Zheng, H., & Wang, F. (2023). Real-Time target detection system for intelligent vehicles based on multi-source data fusion. Sensors, 23(4), 1823. https://doi.org/10.3390/s23041823

[18] Ni, Y. S., Chen, W. L., Liu, Y., Wu, M. H., & Guo, J. I. (2024). Optimizing automated optical inspection: an adaptive fusion and semi-supervised self-learning approach for elevated accuracy and efficiency in scenarios with scarce labeled data. Sensors, 24(17), 5737. https://doi.org/10.3390/s24175737

[19] Bocu, R., Bocu, D., & Iavich, M. (2021). Objects detection using sensors data fusion in autonomous driving scenarios. Electronics, 10(23), 2903. https://doi.org/10.3390/electronics10232903

[20] Xu, M., Qi, J., Wang, X., Zhou, M., Yang, Y., & Yang, P. (2024, December). Leveraging multi-sensor data and domain adaptation for improved Parkinson's disease assessment. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 6016-6021). IEEE. https://doi.org/10.1109/BIBM62325.2024.10822598

[21] Xu, H., Tang, W., Li, Z., Qin, K., & Zou, J. (2024). Multimodal dual cross-attention fusion strategy for autonomous garbage classification system. IEEE Transactions on Industrial Informatics, 20(11), 13319-13329. https://doi.org/10.1109/TII.2024.3435508

[22] Zhou, F., Tao, C., Gao, Z., Zhang, Z., Zheng, S., & Zhu, Y. (2023). 3-D dynamic multitarget detection algorithm based on cross-view feature fusion. IEEE Transactions on Artificial Intelligence, 5(6), 3146-3159. https://doi.org/10.1109/TAI.2023.3342104

[23] Shahian Jahromi, B., Tulabandhula, T., & Cetin, S. (2019). Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles. Sensors, 19(20), 4357. https://doi.org/10.3390/s19204357

[24] Kashinath, S. A., Mostafa, S. A., Mustapha, A., Mahdin, H., Lim, D., Mahmoud, M. A., ... & Yang, T. J. (2021). Review of data fusion methods for real-time and multi-sensor traffic flow analysis. IEEE Access, 9, 51258-51276. https://doi.org/10.1109/ACCESS.2021.3069770

[25] Zhou, B., Liu, J., Cui, S., & Zhao, Y. (2024). A large-scale spatio-temporal multimodal fusion framework for traffic prediction. Big Data Mining and Analytics, 7(3), 621-636. https://doi.org/10.26599/BDMA.2024.9020020

[26] Nai, K., Li, Z., & Wang, H. (2022). Dynamic feature fusion with spatial-temporal context for robust object tracking. Pattern Recognition, 130, 108775. https://doi.org/10.1016/j.patcog.2022.108775

[27] Li, Q., Xu, P., He, D., Wu, Y., Tan, H., & Yang, X. (2024). Multi-source information fusion graph convolution network for traffic flow prediction. Expert Systems with Applications, 252, 124288. https://doi.org/10.1016/j.eswa.2024.124288

[28] Lu, X., Zhong, Y., & Zhang, L. (2022). Open-source data-driven cross-domain road detection from very high-resolution remote sensing imagery. IEEE Transactions on Image Processing, 31, 6847-6862. https://doi.org/10.1109/TIP.2022.3216481

[29] Liu, C., Zhang, S., Hu, M., & Song, Q. (2024). Object detection in remote sensing images based on adaptive multi-scale feature fusion method. Remote Sensing, 16(5), 907. https://doi.org/10.3390/rs16050907

[30] Sun, R., & Ren, Y. (2024). A multi-source heterogeneous data fusion method for intelligent systems in the Internet of Things. Intelligent systems with applications, 23, 200424. https://doi.org/10.1016/j.iswa.2024.200424

[31] Balamuralidhar, N., Tilon, S., & Nex, F. (2021). MultEYE: Monitoring system for real-time vehicle detection, tracking and speed estimation from UAV imagery on edge-computing platforms. Remote sensing, 13(4), 573. https://doi.org/10.3390/rs13040573

[32] Zhao, J., Jia, Y., Ma, L., & Yu, L. (2024). Recurrent adaptive graph reasoning network with region and boundary interaction for salient object detection in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-20. https://doi.org/10.1109/TGRS.2024.3421950

[33] Wang, S., & Ahmad, N. S. (2024). A comprehensive review on sensor fusion techniques for localization of a dynamic target in GPS-denied environments. IEEE Access, 13, 2252-2285. https://doi.org/10.1109/ACCESS.2024.3519874

[34] Zhu, Y., Liang, S., Gong, M., & Yan, J. (2022). Decomposed POMDP optimization-based sensor management for multi-target tracking in passive multi-sensor systems. IEEE Sensors Journal, 22(4), 3565-3578. https://doi.org/10.1109/JSEN.2021.3139365

[35] Su, Y., Chen, X., Liu, G., Cang, C., & Rao, P. (2023). Implementation of real-time space target detection and tracking algorithm for space-based surveillance. Remote Sensing, 15(12), 3156. https://doi.org/10.3390/rs15123156