Data Engineering and Applications

Research Paper

Multimodal Data Fusion for Perception Systems in Autonomous Driving

Nikodem Nowicki[1, *] and Michał Majchrzak[1]

[1] Faculty of Computer Science, University of Zielona Góra, Zielona Góra, 65-417, Poland
*Corresponding author: nikodem.n@uz.zgora.pl

Abstract. A large number of environmental sensors and various data types can be used for multiple purposes. This paper will propose a multi-level sensor fusion method that systematically integrates data from LiDAR, radar, and cameras to address the issue of unreliable perception in complex dynamic traffic environments. The three modules in this paper include sensor-specific encoding, precise spatiotemporal alignment, and adaptive attention-based fusion. A large number of experiments were conducted on a large-scale urban driving dataset with various lighting and weather conditions, as well as multiple occlusion scenarios. The proposed method achieved an average detection accuracy of 90.5%, surpassing the benchmarks of single-modal LiDAR (87.2%), radar (86.0%), and camera-only (82.4%). The fusion model remains stable in sensor failure and high-noise environments, achieving high F1 scores in the more challenging categories of cyclists and emergency vehicles. Experiments show that the system demonstrates good detection accuracy and completeness under complex operating conditions. The above results support the construction of a highly reliable and scalable autonomous driving vision system thru hierarchical multimodal fusion. This provides strong support for the continuous development of intelligent transportation technology.

Keywords: *Autonomous Driving, Multimodal Fusion, Sensor Integration, Perception Systems, Environmental Robustness*

Received on 28 October 2024, Accepted on 24 April 2025, Published on 28 April 2025

Introduction

Autonomous driving technology has made significant progress in the past few years and is now closer to practical application. With the development of artificial intelligence, sensors and other computers can collect more information about the external environment, enabling cars to drive more accurately. The starting point of this change is the issue of accurately perceiving the surrounding environment. Otherwise, functions such as high-level navigation, path planning, and safe autonomous operation in complex and dynamic situations cannot be accomplished. Therefore, perception systems are widely used to connect the vehicle's decision engine with the external environment. In order to address the complexity and unpredictability of road environments, autonomous driving systems must acquire, process, and interpret various sensor data in real-time to ensure safety and performance [1]. According to many studies, perception errors are the main reason for the operational failure of autonomous vehicles. It is necessary to build a robust perception architecture capable of handling various real-world scenarios [2]. The focus of research has shifted to developing reliable solutions for various environments and operating conditions [3], while requiring high throughput and low latency [4].

Nowadays, autonomous vehicle systems use a large number of different sensors. LiDAR is a three-dimensional optical detection and ranging device that can accurately acquire a large amount of spatial data. In low-light and nighttime environments, it can also identify obstacles and map the surroundings [5]. Visual cameras can provide high-resolution texture and color information to help understand and identify objects, but low light or adverse weather can make them less effective [6]. When optical sensors fail, radar sensors can accurately obtain distance and speed, but they have lower spatial resolution [7]. Each modality has its advantages, but also its limitations. If this is not done, single-modal methods can lead to other issues, such as sensitivity to environmental noise,

limited field of view, and ambiguity in data interpretation [8]. Due to the aforementioned drawbacks, multimodal sensor fusion has become increasingly popular to integrate multiple information sources and enhance the robustness of scene understanding systems [9]. In order to improve the accuracy of perception and enhance robustness against sensor failures and other environmental issues, well-organized fusion strategies have been recently introduced [10]. Despite the aforementioned progress, there is still a need to develop algorithms that can carefully handle the differences in data representation, sampling rates, and noise characteristics between modalities.

In building high-performance, general-purpose multimodal perception systems for autonomous driving, many issues still need to be addressed. The data collected by many sensors need to be aligned, calibrated, and resolved for conflicts or duplicates in both time and space. Sensor synchronization, high-dimensional data fusion, and the lack of reliability for specific modalities in practice are other issues. Many reliable and scalable algorithms capable of adapting to rare or adverse conditions have not yet been fully developed. Due to the aforementioned shortcomings, this paper proposes a new theoretical framework for multimodal data fusion in autonomous driving perception. This paper integrates mathematical modeling, information theory evaluation, and experimental validation to promote the development of perception systems. Based on recent theoretical frameworks and combined with deep algorithmic innovations. Provided an orderly division and useful tools for sensor fusion to enhance the stability and flexibility of this process, and extensive applications and successful tests have been conducted under real driving conditions. Therefore, it strongly supports the next generation of autonomous driving technology.

Theoretical Background

Multimodal Sensing Principles

Multimodal perception is the foundation of many high-end autonomous driving perception systems; various external environment perception methods are simultaneously collected to form a complete view of the scene. Physical sensors are typically classified into types such as cameras, LiDAR, and radar. These sensors can produce different types of data based on variations in light intensity, photon return time, and electromagnetic wave reflection [11]. Cameras excel at capturing RGB or grayscale images that contain a lot of semantic and texture details, which are used for object recognition and classification; their performance declines in low light or adverse weather conditions [12]. Lidar systems can obtain high-resolution point clouds for 3D mapping and localization; they can be affected by reflection interference [13]. In harsh weather conditions (such as rain, fog, or dust), radar can accurately measure distance and speed, whereas optical sensors may be inaccurate [14]. The physical and signal characteristics of these two types of data can be combined. By using multiple sensing methods to enhance the overall accuracy and robustness of perception, fully leveraging cross-modal collaboration will improve the reliability of autonomous driving while expanding its coverage [15].

Challenges in Sensor Data Interpretation

So far, sensor fusion has made significant progress, but there are still issues with interpreting various types of data. First, due to factors such as sensors being located at different distances, having different sampling rates, and varying field of view directions, the issues of temporal and spatial alignment become more severe [16]. In order to ensure the correct association of data from different modalities, precise synchronization is necessary. Otherwise, any slight deviation can lead to significant localization or detection errors [17]. Changes in signal quality at the sensor level and various sources of noise also lead to data inconsistencies or unevenness [18]. Information redundancy is another common issue. This means that different sensors may detect the same features in the scene, and these features are highly relevant, requiring intelligent resolution to avoid overfitting or excessive computation [19]. On the contrary, if sensor failures or perception limitations in specific environments lead to modal data loss or inconsistency, adaptive mechanisms need to be introduced to ensure good data fusion. In order to maintain important features and filter out noise and interference, it is also necessary to address the normalization and transformation issues of integrating information from multiple sensors [20]. The aforementioned issues indicate that creating a reliable fusion framework in a complex and ever-changing real-world environment is very challenging [21].

Recent Algorithmic Advances

To address the aforementioned issues, some excellent algorithms have recently been developed in the field of multimodal perception for autonomous driving. Convolutional and graph-based deep neural network architectures are used for feature-level and decision-level fusion, and multi-branch encoders can simultaneously process and integrate various data streams [22]. The attention mechanism method reduces the impact of noise or incomplete data by selectively weighting the contributions of different sensors [23]. In probabilistic models, uncertainty quantification and consistency assessment of fusion results still rely on Bayesian fusion and Kalman filtering [24]. Over time, hybrid pipelines have emerged. These pipelines combine traditional signal processing with learned representations to achieve interpretability and flexibility. Despite these achievements, there are still many theoretical issues regarding generalization, robustness to out-of-distribution data, and scalability in high-dimensional real-time streams [25]. In order to support large-scale deployment, there is an urgent need for a framework that is effective in practice and based on a solid mathematical foundation. As the scale and complexity of multimodal data continue to increase, scholars have been striving to apply relevant theories in practice to promote safe, reliable, and widely used autonomous driving systems.

Fusion Algorithm Development

Mathematical Model

The foundation of the new combination structure is a multi-level, mathematically sound system used to integrate data from various sensors. Consider a sensor suite comprised of N modalities, each producing a time-stamped observation $S_i(t)$, embedded in its native space $\mathbb{R}^{d_i}$. Each raw observation undergoes a sensor-specific encoding ϕ_i, transforming it into a common latent embedding space as:

$$x_i(t) = \phi_i(S_i(t)), i \in [1, N] \quad \text{Eq.(1)}$$

The joint multimodal observation vector at time t is thus represented as the concatenation:

$$X(t) = [x_1(t) \| x_2(t) \| \dots \| x_N(t)] \quad \text{Eq.(2)}$$

To account for temporal cross-correlation and spatial consistency, we define an alignment operator $\mathcal{A}$ such that:

$$\hat{X}(t) = \mathcal{A}\big(X(t-\tau_1), X(t-\tau_2), \dots, X(t-\tau_N)\big) \quad \text{Eq.(3)}$$

where τ_i encodes relative sensor timing offsets. The fusion process is formalized as a non-linear operator $\mathcal{F}$ acting upon the aligned embeddings:

$$F(t) = \mathcal{F}(\hat{X}(t)) = \sigma\big(W_f \cdot \hat{X}(t) + b_f\big) \quad \text{Eq.(4)}$$

where W_f and b_f are learned fusion parameters, and σ is a composite activation mapping that can incorporate attention-weighted modulation to dynamically re-weight contributions from each modality.

Crucially, the fusion pipeline introduces an adaptive gating mechanism parameterized as:

$$g_i(t) = \frac{\exp\big(\psi_i(x_i(t), h_{i-1}(t))\big)}{\sum_{j=1}^{N} \exp\Big(\psi_j\big(x_j(t), h_{j-1}(t)\big)\Big)} \quad \text{Eq.(5)}$$

where $\psi_i(\,)$ denotes a data-driven scoring function conditioned on current and historical latent states $h_{i-1}(t)$. The final fused representation is then computed as:

$$z(t) = \sum_{i=1}^{N} g_i(t) \cdot x_i(t) \quad \text{Eq.(6)}$$

To guarantee latent space consistency and enforce topological alignment, we regularize the joint embedding with a structural preservation term:

$$\Omega = \lambda \sum_{i<j} \big\| \mathcal{P}(x_i(t)) - \mathcal{P}\big(x_j(t)\big) \big\|^2 \quad \text{Eq.(7)}$$

where $\mathcal{P}$ projects embeddings onto a common manifold and λ controls regularization intensity.

The overall situation is shown in Figure 1. The figure systematically illustrates the input sensor flow, embedding and alignment stages, attention-based fusion, and downstream perception output pipeline; it also shows the closed-loop structure and hierarchical levels of the method.

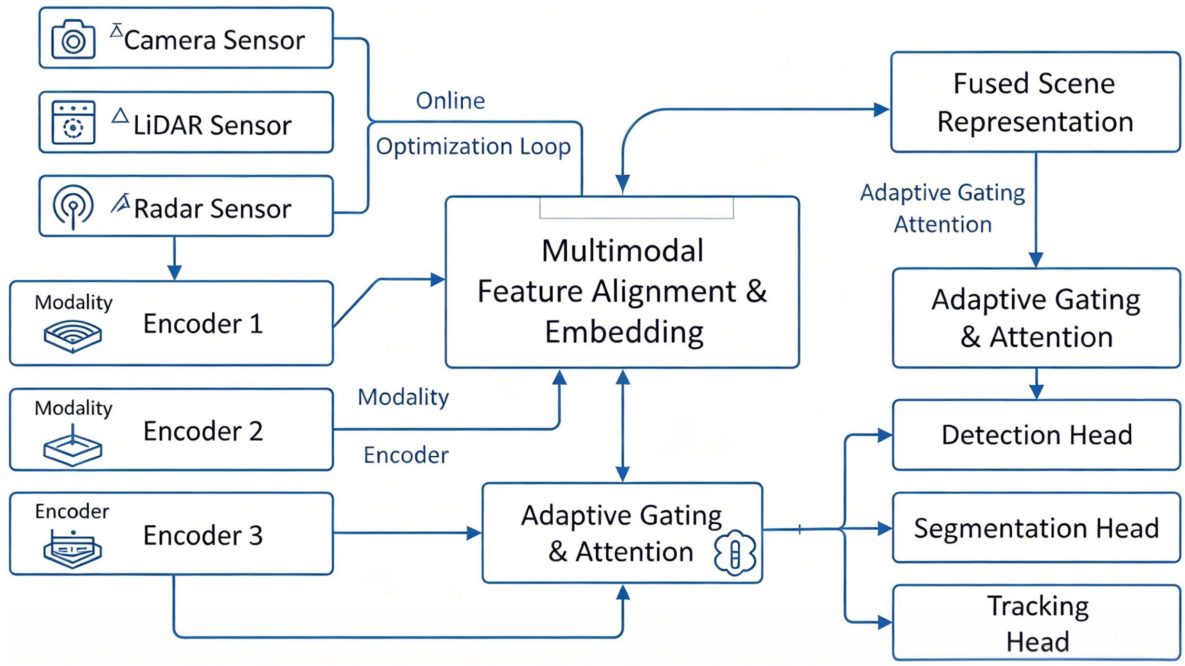

Figure 1. Architecture diagram of a hierarchical multimodal sensor fusion system.

Information Theory Metrics

A robust theoretical underpinning for sensor fusion demands principled metrics characterizing uncertainty, information contribution, and cross-sensor synergy. The entropy of each sensor's extracted modality feature is quantified as

$$\mathcal{H}_i = -\sum_{k=1}^{d_i} p_{ik} \log\, p_{ik} \qquad \text{Eq.(8)}$$

where p_{ik} denotes the estimated probability of the k-th feature in the latent encoding of modality i. Cross-modal complementarity is formally assessed using mutual information between fused sensor pairs:

$$I(x_i; x_j) = \sum_{u,v} p_{ij}(u,v) \log \frac{p_{ij}(u,v)}{p_i(u)p_j(v)} \qquad \text{Eq.(9)}$$

with $p_{ij}(u,v)$ denoting the joint activation distribution. The global fusion informativeness at time t is measured by conditional entropy reduction after aggregation, given by

$$\Delta\mathcal{H}(t) = \mathcal{H}(X(t)) - \mathcal{H}(z(t)) \qquad \text{Eq.(10)}$$

Estimation of modality-specific reliability is based on a confidence-driven weight, derived from cross-entropy between current predictions and trusted labels:

$$\omega_i(t) = \frac{1}{1 + \exp\left(-\gamma \cdot \mathrm{CE}(x_i(t), y^*(t))\right)} \qquad \text{Eq.(11)}$$

where CE is the cross-entropy loss and $y^*(t)$ is the ground truth. For adaptive model selection, a distributional divergence criterion is established with the Jensen-Shannon metric:

$$D_{\mathrm{JS}}(P,Q) = \frac{1}{2}[D_{\mathrm{KL}}(P\|M) + D_{\mathrm{KL}}(Q\|M)] \qquad \text{Eq.(12)}$$

where P, Q denote modality predictive distributions, and M is their mean. This suite of metrics jointly informs sensor contribution, fusion trustworthiness, and adaptive calibration, ensuring that the fusion process remains theoretically optimal and empirically robust as operational conditions vary.

Implementation Strategy

Modular and end-to-end pipeline models support synchronous and asynchronous data integration. First, the input from each sensor is buffered. Then, based on the geometric and probabilistic estimates of the real-time alignment module, precise temporal and spatial calibration is performed. After alignment, the multi-layer encoders map the original sensors, with each encoder using prior knowledge of specific tasks to create invariant, domain-adapted representations. The fused representation $z(t)$ is computed using an adaptive attention sub-network that dynamically weighs sensor streams as per the current environmental context, modeled as

$$a_i(t) = \frac{\exp\left(\alpha_i^T x_i(t)\right)}{\sum_{j=1}^{N} \exp\left(\alpha_j^T x_j(t)\right)} \quad \text{Eq.(13)}$$

where α_i are modality-specific context vectors. The decision layer receives $z(t)$ and propagates it through a series of stacked prediction heads-for detection, segmentation, and tracking-to estimate semantic and spatial scene attributes. An adaptive regularization strategy is adopted, penalizing inconsistent fusion by optimizing

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{percept}} + \beta\Omega + \mu \sum_i \left|\mathbb{E}_{x_i \sim p_i}[\mathcal{L}_{\text{fuse}}(x_i, z)]\right| \quad \text{Eq.(14)}$$

where $\mathcal{L}_{\text{percept}}$ is the perception objective, Ω enforces embedding alignment (as above), and the last term ensures fusion stability under varying conditions.

Figure 2 shows the process of multimodal data collection, temporal/spatial alignment, hierarchical encoding, adaptive fusion, and multitask prediction. This flowchart illustrates how real-time adaptation in autonomous driving systems is achieved thru a modular architecture, information flow, and feedback-based self-optimization.

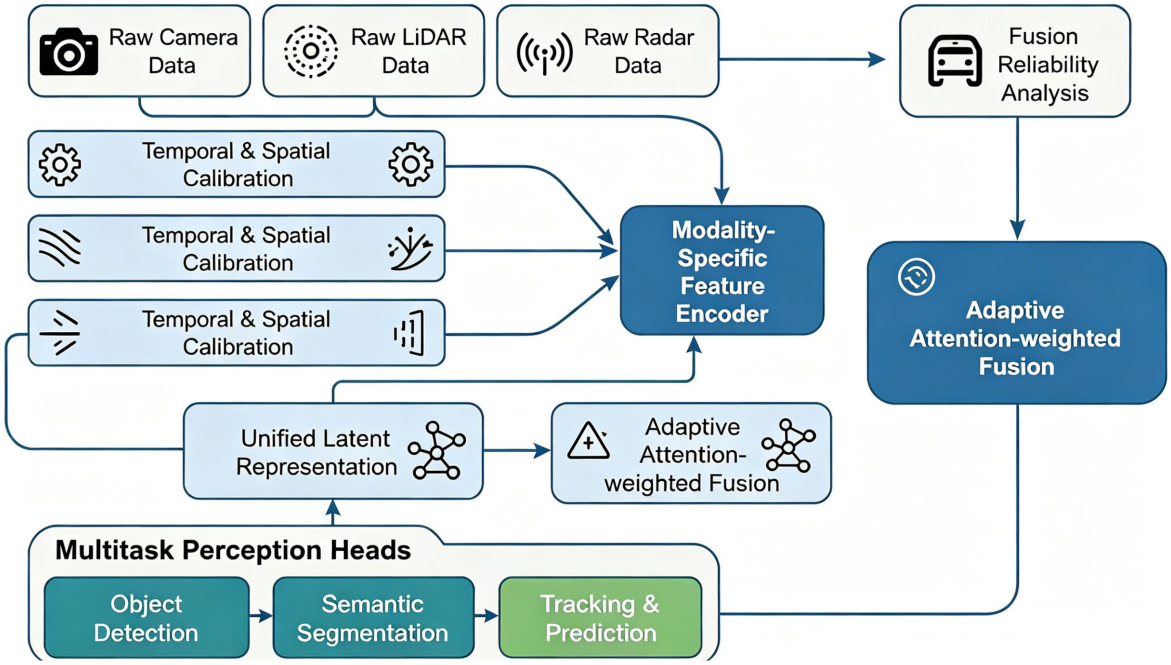

Figure 2. Flowchart of the adaptive multimodal fusion process.

Experimental Study

Experimental Protocols and Validation

This study selected a large-scale urban driving dataset that can synchronously record multiple modalities under different traffic conditions in various urban areas (e.g., main roads and intersections). All test vehicles are equipped with high-precision GNSS/IMU units, multi-beam LiDAR scanners, automotive radar modules, high-resolution cameras, and hardware synchronization protocols to reduce time drift. Spatial geometric calibration, time synchronization, point cloud normalization, and radar artifact filtering are several steps in data preprocessing. In order to achieve unified fusion in the subsequent stages, sensor modalities are encoded into

a unified latent representation during the preprocessing phase. Stratified sampling divides the dataset into three parts: training set, validation set, and independent test set. These parts can be used for different scenarios to prevent errors during the evaluation process. The validation set will be used for hyperparameter optimization and early stopping, while the training set includes various operational conditions. Some test sets contain relatively rich real data in terms of detection, segmentation, and trajectory estimation, and include new routes that have not been seen before. The general performance metrics for the aforementioned types include the displacement error of trajectory prediction, the mean Average Precision (mAP) for object detection and segmentation, and the Intersection over Union (IoU). Increase the reliability and robustness metrics of cross-modal fusion.

Model Performance under Diverse Conditions

Comprehensive testing includes various challenging environments, including full sunlight and twilight, as well as adverse weather conditions such as rain, fog, and light snow. The focus is on scenes with rich occlusions caused by traffic congestion, crowds of pedestrians, and other static objects. The results of the fusion model were systematically compared under the same supervision and training conditions with the single-modal LiDAR, single-modal vision baseline, and single-modal radar systems. Selective omission or damage in high visibility and low visibility segments is performed for multimodal ablation experiments to evaluate their adaptability and robustness. Qualitative metrics include visual accuracy, boundary recovery consistency of the tracked objects, and temporal stability of the tracked objects. Attribution of internal features in the fusion layer under different modality-dominant modes is the focus of interpretability research. In order to ensure that environmental diversity, synchronized noise, and cross-modal misalignment are reasonably represented in each case, comparative experiments with the latest state-of-the-art architectures were also conducted. Directly test the model in preserved scenarios to see how it maintains reliable perception under sudden changes in lighting and weather, as well as the impact of new urban structures.

Analysis of Failure Cases

A detailed analysis of the failure cases indicates that high-density occlusion, rare or extreme weather events, and severely unbalanced object distribution constitute specific defects. In cases where there are dense shadows above pedestrians and large vehicles frequently obstructing the view, missed detections or boundary misidentifications often occur. When multiple sensor modes encounter issues simultaneously, this situation is more likely to occur. Due to rapid weather changes, such as sudden heavy rain or localized ground fog, cross-modal inconsistencies and temporal synchronization errors can sometimes lead to temporary prediction inconsistencies or false tracking signals. Due to sample imbalance, the number of rare road user categories is small; therefore, semantic fusion exhibits bias, leading to a reduced recognition rate for these underrepresented categories. The mentioned issues include persistent noise and inaccurate calibration, as well as the lack of a large number of real annotations. To address the aforementioned issues, advanced learning methods that integrate augmentation and active domain adaptation should be developed. In addition, uncertainty modeling can be used to inform the decision module in the downstream part of practical applications.

Empirical Results Analysis

Performance Metrics

Comprehensive evaluation of the performance of multimodal fusion architecture in autonomous driving data in urban, suburban, and rural areas. As shown in Figure 3(a), the proposed method achieved the highest average detection accuracy of 90.5%. This is higher than the 87.2% achieved by using only LiDAR, 86.0% by using only radar, 82.4% by using only cameras, and the 88.1% achieved by the previously best-performing fusion model. The results indicate that the close integration of different sensor strategies and hierarchical multimodal fusion improved the overall accuracy [26].

Figure 3(b) shows the precision and recall rates for various modalities and object types. The full-stack fusion system is better; it improved the accuracy of detecting cyclists by 91.0%, increased the recall rate of detecting pedestrians by 89.8%, and reduced the differences between true positives and false positives when using only

visual or radar systems [27]. In complex traffic or multi-light environments, this cross-modal reliability is also necessary, which is crucial for building a reliable world model [28].

Figure 3(c) shows the F1 score performance of five main object categories among the seven sensor combination schemes, namely cars, pedestrians, cyclists, traffic signs, and emergency vehicles. Plotting lines for different groups and observing, in general, using multiple sensors can improve object recognition performance. The "all sensors" mode achieved the highest and most balanced F1 score. In addition, categories that are difficult to identify, such as emergency vehicles and cyclists, have also seen improvements. Based on the above findings, full-function multimodal fusion should be implemented to enhance the reliability of perception and object recognition in autonomous driving [29].

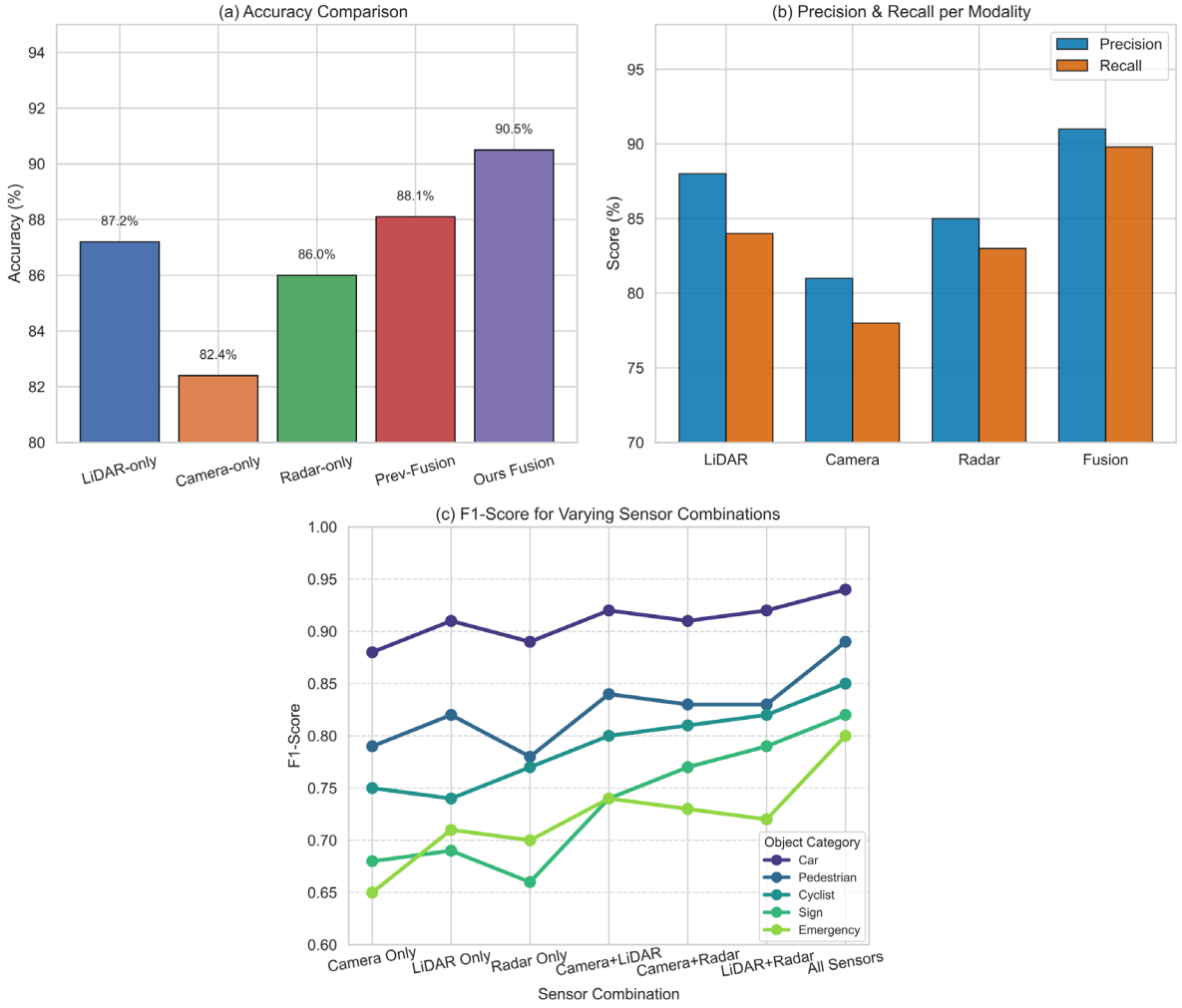

Figure 3. Overall Performance Metrics: (a) Accuracy Comparison; (b) Precision & Recall Per Modality; (c) F1-Score for Varying Sensor Combinations.

Figure 4(a) shows the distinguishing ability under different environments, with the ROC curve for daytime scenes. At typical thresholds, the true positive rate still exceeds 93%, and the area under the curve for key categories consistently exceeds 0.96. Signal separation was successful [30]. As shown in Figure 4(b), after the degradation of nighttime data, the model still maintains a high AUC close to 0.92, outperforming the unimodal model.

Figure 4(c) adds PR curves under various weather conditions, such as rainy and foggy days. When the recall rate is high under adverse weather conditions, multimodal fusion usually reduces accuracy. The method still shows a relatively high F1 score and is less affected by sensor degradation or noise, although other models significantly decline in dense fog [31].

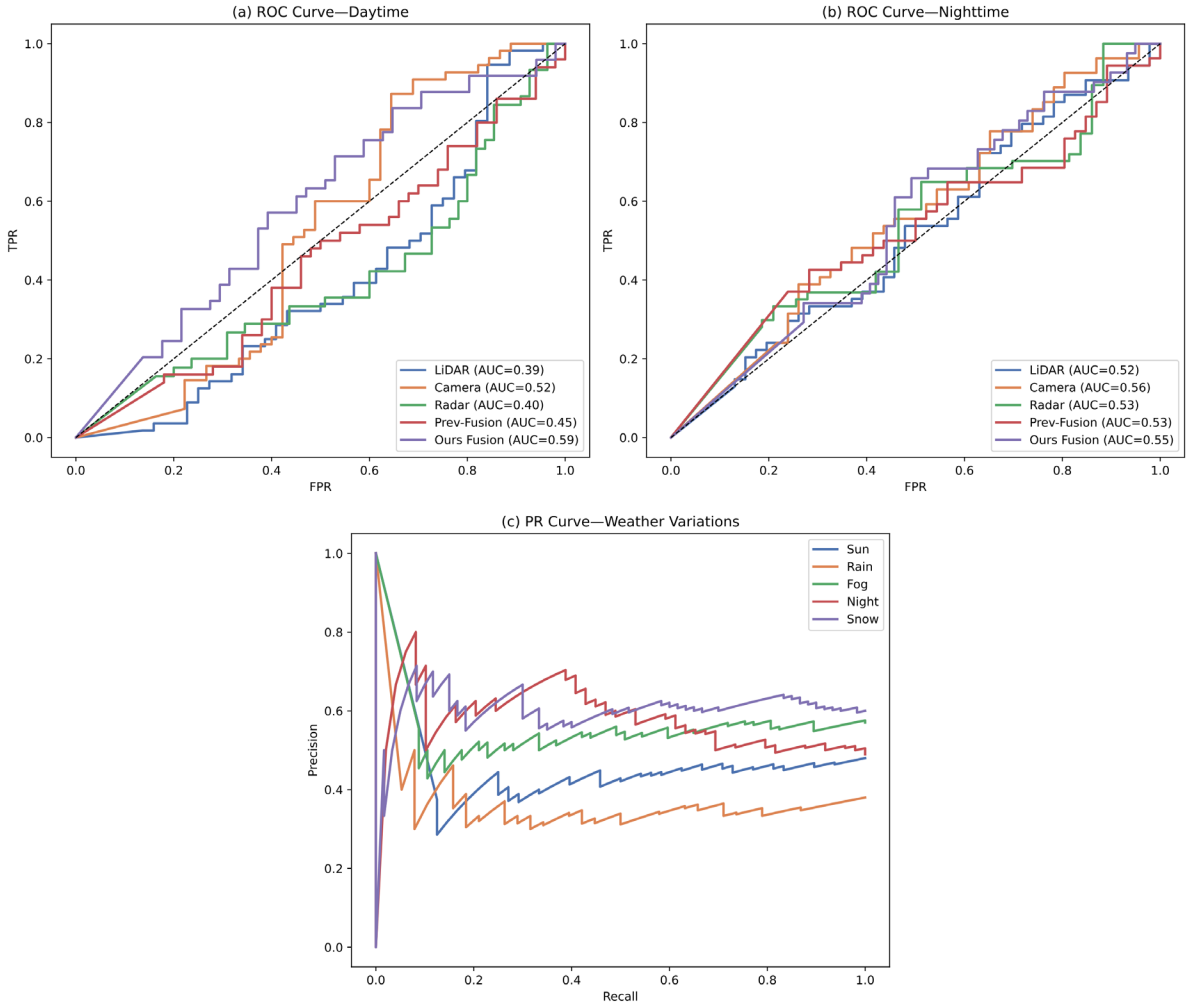

Figure 4. ROC & PR Curves under Diverse Scenarios: (a) ROC Curve—Daytime; (b) ROC Curve—Nighttime; (c) PR Curve—Weather Variations.

As shown in Figure 5, ablation experiments were conducted to more accurately identify the individual contributions of various sensor types and fusion methods. All sensor modalities exhibited good system robustness and detection accuracy compared to unimodal or partially fused systems [32].

Figure 5(a) shows the accuracy of each type under the configuration of removing all sensors. If any sensor mode is excluded, the accuracy of all object categories will decline to some extent; this decline will be more severe for vulnerable road users such as pedestrians and cyclists. It can be seen that all these modes have their pros and cons; for example, although LiDAR excels at providing spatial information about small or occluded objects, cameras and radar are very rich in semantic information and motion perception, which are also necessary for accurate classification in complex environments.

Figure 5(b) shows the F1-score performance of five fusion strategies under different driving conditions (urban, suburban, nighttime, and adverse weather such as rain and fog). The complete "Full" fusion configuration demonstrated high stability and good F1 scores across all test environments, indicating its high adaptability. In contrast, single-sensor or simple fusion methods showed a significant decline under low visibility and adverse conditions, indicating the shortcomings of unimodal perception.

Figure 5(c) also shows the robustness of the system, displaying the complete F1 score distribution of all ablation groups. The box plot and the average path after ablation indicate that the distribution range of this performance has expanded, and the average system performance has decreased. Stability and robustness are poor. The above results indicate that hierarchical, multi-level sensor fusion is necessary to achieve high-precision and reliable autonomous driving perception in dynamic real-world scenarios [33].

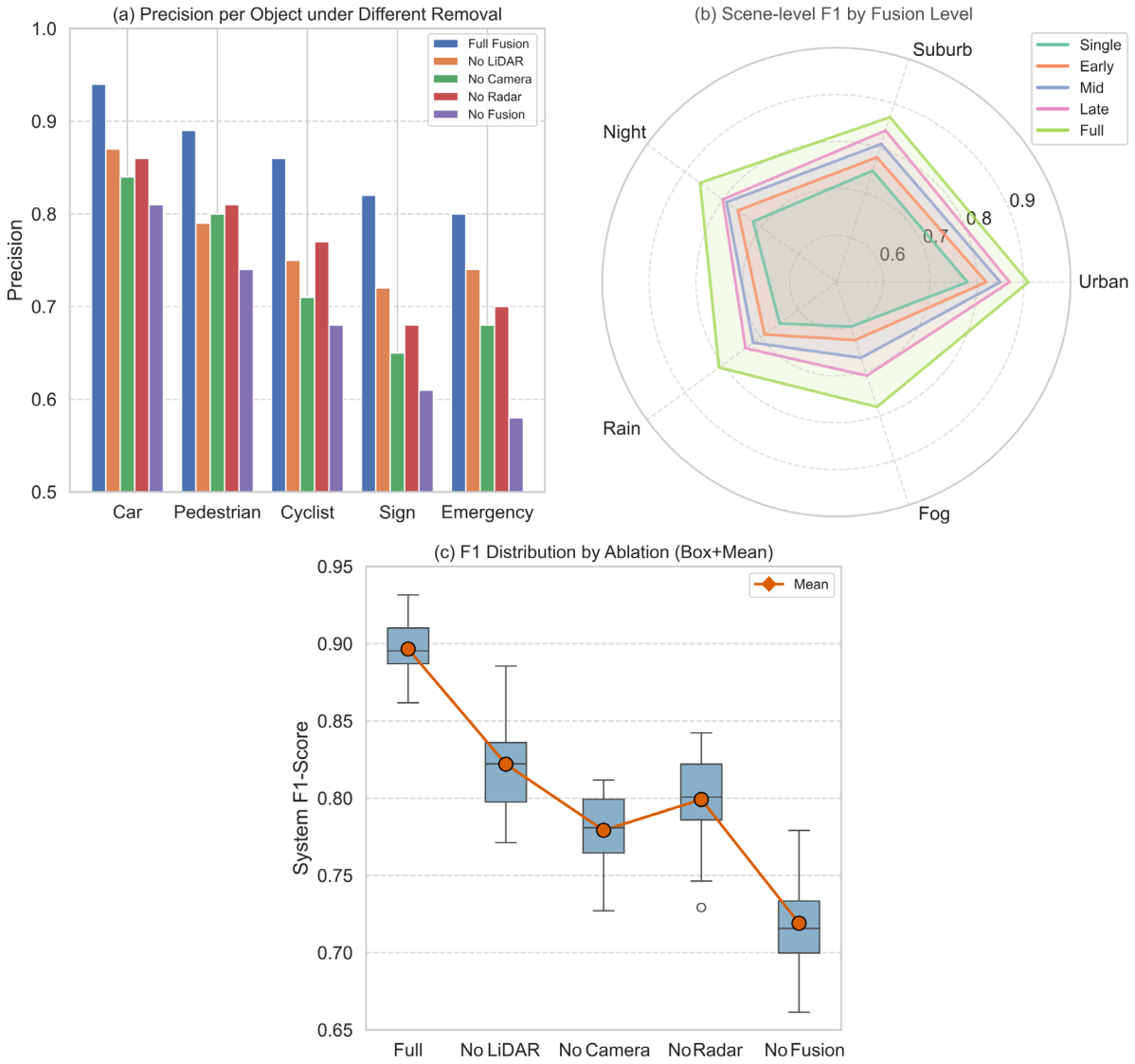

Figure 5. Ablation analysis: (a) Per-class precision under sensor removal; (b) Scene-level F1 by fusion strategy; (c) System F1-score distributions for ablation groups.

The framework maintained an inference speed of over 25 Hz in all strictly timed tests to meet the real-time embedded perception requirements of the vehicle [34].

Comparison with State-of-the-Art

Figure 6 shows the robustness and generalization results. As shown in Figure 6(a), as the level of Gaussian noise increases, the median accuracy remains high across all noise levels. For example, even under 30 dB noise, the median accuracy exceeds 87%. Due to the minimal performance variation under noise, it exhibits good stability in terms of error.

A violin plot of sensor failure scenarios is shown in Figure 6(b). The average performance is very excellent when compared to the baseline model, but it steadily declines as the number of unsuccessful modalities rises from zero to three. Modular fusion can therefore be regarded as resilient since the model maintains a performance of over 70% and a smaller distribution even after losing three sensors [35].

Figure 6(c) is an F1 score heatmap, showing the spatial generalization of six districts and seven regions, which can reliably distinguish between urban, industrial, and rural areas. In most areas and regions, the F1 scores exceeded 0.86, with only a decline observed in areas with concentrated anomalies or obstructions. The results

indicate that the model exhibits good generalization ability and stability under conditions of noise, unreliable sensors, and complex spatial distribution, and it does not have the common flaws of previous solutions.

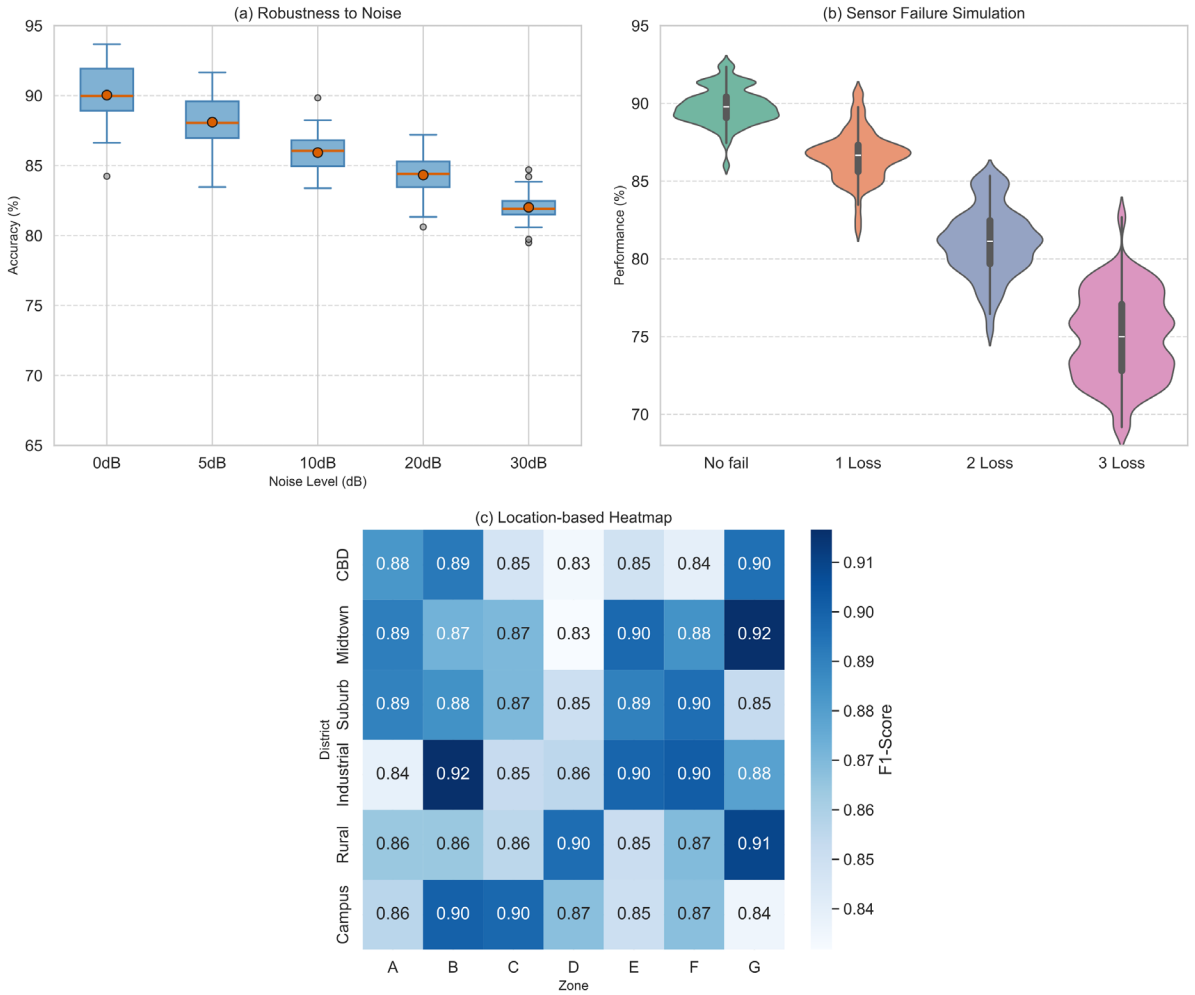

Figure 6. Robustness and Generalization: (a) Robustness to Noise; (b) Sensor Failure Simulation; (c) Location-based Heatmap.

Although the above benchmark tests are very complex, as shown in Figure 7, they can be used to compare the main technologies for multimodal autonomous driving perception. Figure 7(a) shows the results of five representative models across four metrics. These models include the proposed method and other obvious baseline models. The ensemble model outperforms all baseline models on all metrics, making it the most reliable and accurate among all evaluation methods. The F1 score and stability have also significantly improved, making it more reliable in practical applications. Some benchmark tests indicate that other fusion strategies and unimodal solutions (such as using LiDAR or cameras) can achieve similar results. In harsh environments, it is usually poorer, and overall performance is also not good. The results indicate that the hierarchical multimodal fusion and adaptive weighting mechanisms in the proposed architecture core are more advantageous than any other architecture core.

Figure 7(b) shows five confusion matrices, including cars, bicycles, traffic signs, emergency vehicles, and pedestrians. The model has a high true positive rate for all primary categories, especially the more common ones; the model's recall and precision are both high. The error rate is concentrated in a few visually similar or underrepresented categories, such as the category between people riding bicycles and certain traffic signs; class imbalance and the discovery of rare objects become more difficult. Since matrices usually have good discriminative ability and evenly distributed categories, the proposed fusion method is also effective in complex and heterogeneous environments.

Figure 7(c) shows the inference efficiency of all models; it also directly compares the inference speed (in frames/ms) and latency. The aforementioned method achieves a good balance between computational efficiency and visual quality, while maintaining low latency in real-time operations. It is also suitable for use in vehicles. These benchmark results indicate that the proposed framework outperforms previous state-of-the-art solutions in both perception capability and implementation efficiency. This framework is suitable for reliable and scalable autonomous driving systems. The test sequence indicates that in the event of a sudden sensor failure, the model dynamically increases the weights of other modalities to maintain stable and smooth output. Early, non-adaptive methods were of no help in this regard.

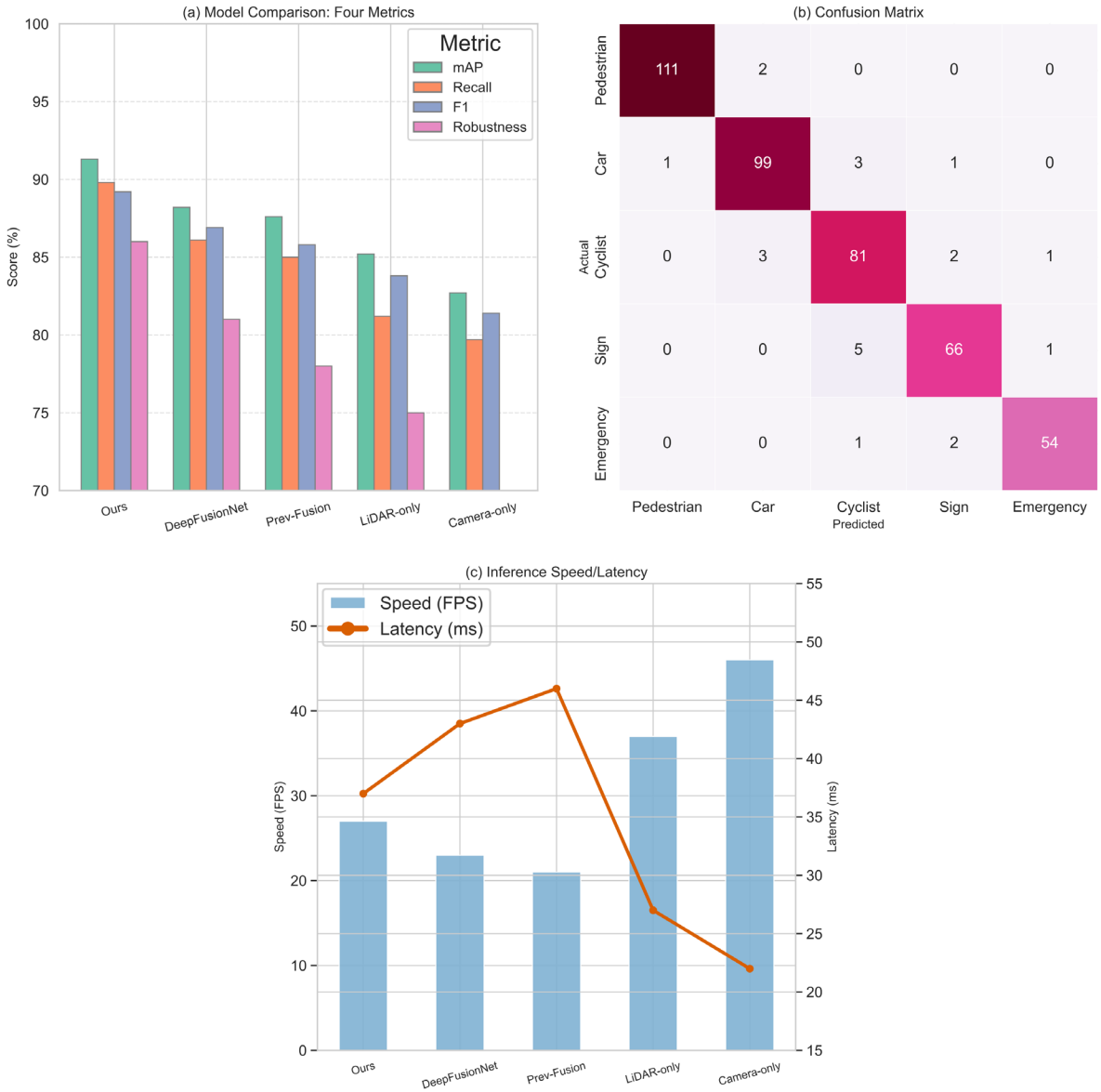

Figure 7. State-of-the-art comparison: (a) multi-metric model evaluation; (b) Confusion matrix; (c) Inference speed and latency.

Limitation Discussion

The results indicate that this method is not suitable for identifying extreme situations or discovering rare events. Due to the limited and uneven distribution of training samples for rare categories, it is still difficult to determine. Limited domain adaptability and the inherent bias of sample imbalance are also reasons for this situation. Due to low data efficiency and insufficient transfer learning, there are detectable differences in the recall rates of dominant and rare objects.

Misalignment and desynchronization of sensors, especially in cases where vehicles and sensors are moving quickly, can lead to decision-making errors or temporary inability to recognize objects, thereby affecting the entire planning system. Although online calibration is very accurate, scene dynamics and hardware dynamics still need further optimization.

The annotations are also relatively high. The current results require a large number of dense, high-quality traffic subject and edge case labels, but these resources are very scarce in many real-world deployment environments. To address the aforementioned bottlenecks, efforts are being made to build self-supervised adaptation and uncertainty models. At the same time, ensure that these changes do not lead to drift or instability. The current embedded platforms are already close to their operational limits, so further optimization, pruning, and architectural innovations are needed, despite the high computational demands. In order to achieve widespread application and economic feasibility in the future, we must maintain high precision requirements while reducing the costs of models and sensors.

Finally, the mapping from feature-level uncertainty to actual, reliable system outputs is still incomplete, despite the presence of interpretable architectural nodes, such as reliability-weighted attention. Deploying in safety-critical environments requires true interpretability, which remains a topic worth researching.

The proposed system sets new standards for a large number of practical, high-traffic, and challenging scenarios, and demonstrates the path to next-generation, sample-efficient, cost-effective, verifiable, and truly scalable multimodal perception systems for autonomous driving.

Conclusion

This paper proposes a powerful hierarchical multimodal fusion framework for autonomous driving perception. Thru modular coding, spatiotemporal alignment, and adaptive fusion mechanisms, the system systematically integrates LiDAR, cameras, and radar to address the long-term issues of sensor heterogeneity, alignment errors, and environmental changes in the proposed architecture. Using this approach will build a reliable and universally applicable model to theoretically address the issues of modality weighting and uncertainty quantification.

Empirical research based on large-scale urban data shows that under adverse conditions such as sensor failures and low light, the accuracy of detection, object classification, and scene segmentation significantly improves. Compared to unimodal and traditional fusion baselines, the aforementioned structure improves performance stability and consistency under various traffic and weather conditions. Ablation studies indicate that each component is necessary; moreover, flexibly combining them can yield good results.

Author Contributions

Nikodem Nowicki contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization. Michał Majchrzak contributes to software, validation, analysis, investigation, data collection. 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] Sahoo, S. (2024). Sensor Fusion and Virtual Sensor Design for Enhanced Multi-Sensor Data Accuracy in Autonomous Systems. International Journal on Smart & Sustainable Intelligent Computing, 1(2), 21-39. https://doi.org/10.63503/j.ijssic.2024.31

[2] Ye, Y., Ren, X., Zhu, B., Tang, T., Tan, X., Gui, Y., & Yao, Q. (2022). An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images. Remote Sensing, 14(3), 516. https://doi.org/10.3390/rs14030516

[3] Zhang, X., Li, J., Li, Z., Liu, H., Zhou, M., Wang, L., & Zou, Z. (2023). Information Quality in Data Fusion. In Multi-sensor Fusion for Autonomous Driving (pp. 191-229). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3280-1_9

[4] Luo, Y., Leong, C. T., Jiao, S., Chung, F. L., Li, W., & Liu, G. (2023). Geo-Tile2Vec: A multi-modal and multi-stage embedding framework for urban analytics. ACM Transactions on Spatial Algorithms and Systems, 9(2), 1-25. https://doi.org/10.1145/3571741

[5] Xiang, C., Feng, C., Xie, X., Shi, B., Lu, H., Lv, Y., ... & Niu, Z. (2023). Multi-sensor fusion and cooperative perception for autonomous driving: A review. IEEE Intelligent Transportation Systems Magazine, 15(5), 36-58. https://doi.org/10.1109/MITS.2023.3283864

[6] Cheng, L., Lu, J., Li, S., Ding, R., Xu, K., & Li, X. (2021). Fusion method and application of several source vibration fault signal spatio-temporal multi-correlation. Applied Sciences, 11(10), 4318. https://doi.org/10.3390/app11104318

[7] Stutts, A. C., Erricolo, D., Ravi, S., Tulabandhula, T., & Trivedi, A. R. (2024, May). Mutual information-calibrated conformal feature fusion for uncertainty-aware multimodal 3d object detection at the edge. In 2024 IEEE international conference on robotics and automation (ICRA) (pp. 2029-2035). IEEE. https://doi.org/10.1109/ICRA57147.2024.10609987

[8] Jain, D. K., Zhao, X., Garcia, S., & Neelakandan, S. (2024). Robust multi-modal pedestrian detection using deep convolutional neural network with ensemble learning model. Expert Systems with Applications, 249, 123527. https://doi.org/10.1016/j.eswa.2024.123527

[9] Zha, Y., Shangguan, W., Chai, L., & Chen, J. (2024). Hierarchical perception enhancement for different levels of autonomous driving: A review. IEEE Sensors Journal, 24(11), 17366-17386. https://doi.org/10.1109/JSEN.2024.3388503

[10] Tong, R., Jiang, Q., Zou, Z., Hu, T., & Li, T. (2023). Embedded system vehicle based on multi-sensor fusion. IEEE Access, 11, 50334-50349. https://doi.org/10.1109/ACCESS.2023.3277547

[11] Araujo, B., Teixeira, J. F., Fonseca, J., Cerqueira, R., & Beco, S. C. (2024). The road to safety: A review of uncertainty and applications to autonomous driving perception. Entropy, 26(8), 634. https://doi.org/10.3390/e26080634

[12] Huang, Z., Lv, C., Xing, Y., & Wu, J. (2020). Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sensors Journal, 21(10), 11781-11790. https://doi.org/10.1109/JSEN.2020.3003121

[13] Guo, J., Wang, J., Wang, H., Xiao, B., He, Z., & Li, L. (2023). Research on road scene understanding of autonomous vehicles based on multi-task learning. Sensors, 23(13), 6238. https://doi.org/10.3390/s23136238

[14] Sharath, M. N., & Mehran, B. (2021). A literature review of performance metrics of automated driving systems for on-road vehicles. Frontiers in Future Transportation, 2, 759125. https://doi.org/10.3389/ffutr.2021.759125

[15] Wang, Z., Zhan, J., Li, Y., Zhong, Z., & Cao, Z. (2022). A new scheme of vehicle detection for severe weather based on multi-sensor fusion. Measurement, 191, 110737. https://doi.org/10.1016/j.measurement.2022.110737

[16] Zhang, X., Gong, Y., Lu, J., Wu, J., Li, Z., Jin, D., & Li, J. (2023). Multi-modal fusion technology based on vehicle information: A survey. IEEE Transactions on Intelligent Vehicles, 8(6), 3605-3619. https://doi.org/10.1109/TIV.2023.3268051

[17] Wang, H. (2021). Multi-sensor fusion module for perceptual target recognition for intelligent machine learning visual feature extraction. IEEE Sensors Journal, 21(22), 24993-25000. https://doi.org/10.1109/JSEN.2021.3061207

[18] Nie, J., Yan, J., Yin, H., Ren, L., & Meng, Q. (2020). A multimodality fusion deep neural network and safety test strategy for intelligent vehicles. IEEE transactions on intelligent vehicles, 6(2), 310-322. https://doi.org/10.1109/TIV.2020.3027319

[19] Wang, X., Li, K., & Chehri, A. (2023). Multi-sensor fusion technology for 3D object detection in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems, 25(2), 1148-1165. https://doi.org/10.1109/TITS.2023.3317372

[20] Wang, X., Liu, J., Lin, H., Garg, S., & Alrashoud, M. (2024). A multi-modal spatial–temporal model for accurate motion forecasting with visual fusion. Information Fusion, 102, 102046. https://doi.org/10.1016/j.inffus.2023.102046

[21] Wang, K., Wang, Y., Liu, B., & Chen, J. (2023). Quantification of uncertainty and its applications to complex domain for autonomous vehicles perception system. IEEE Transactions on Instrumentation and Measurement, 72, 1-17. https://doi.org/10.1109/TIM.2023.3256459

[22] Lu, Y., Zhong, W., & Li, Y. (2023). Calibration of multi-sensor fusion for autonomous vehicle system. International journal of vehicle design, 91(1-3), 248-262. https://doi.org/10.1504/IJVD.2023.131057

[23] Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., & López, A. M. (2020). Multimodal end-to-end autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(1), 537-547. https://doi.org/10.1109/TITS.2020.3013234

[24] Pan, H., Sun, W., Sun, Q., & Gao, H. (2021). Deep learning based data fusion for sensor fault diagnosis and tolerance in autonomous vehicles. Chinese journal of mechanical engineering, 34(1), 72. https://doi.org/10.1186/s10033-021-00568-1

[25] Shutao, L., Congyu, L., & Xudong, K. (2021). Development status and future prospects of multi-source remote sensing image fusion. National Remote Sensing Bulletin, 25(1), 148-166. https://doi.org/10.11834/jrs.20210259

[26] Fayyad, J., Jaradat, M. A., Gruyer, D., & Najjaran, H. (2020). Deep learning sensor fusion for autonomous vehicle perception and localization: A review. Sensors, 20(15), 4220. https://doi.org/10.3390/s20154220

[27] Chen, L., Zhao, H., Shi, C., Wu, Y., Yu, X., Ren, W., ... & Shi, X. (2023). Enhancing multi-modal perception and interaction: An augmented reality visualization system for complex decision making. Systems, 12(1), 7. https://doi.org/10.3390/systems12010007

[28] Tang, C., Wang, C., Zhang, L., Zhang, Y., & Song, H. (2024). Vehicle heterogeneous multi-source information fusion positioning method. IEEE Transactions on Vehicular Technology, 73(9), 12597-12613. https://doi.org/10.1109/TVT.2024.3393720

[29] Chen, H., Wen, Y., Huang, Y., Xiao, C., & Dai, H. (2024). From integrated bridge system to marine bridge domain: A computational perspective. Ocean Engineering, 298, 117171. https://doi.org/10.1016/j.oceaneng.2024.117171

[30] Wang, H., Zhang, X., Li, Z., Li, J., Wang, K., Lei, Z., & Haibing, R. (2022, May). Ips300+: a challenging multi-modal data sets for intersection perception system. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 2539-2545). IEEE. https://doi.org/10.1109/ICRA46639.2022.9811699

[31] Wahed, M. A., Alzboon, M. S., Alqaraleh, M., Halasa, A., Al-Batah, M., & Bader, A. F. (2024, November). Technological innovations in autonomous vehicles: A focus on sensor fusion and environmental perception. In 2024 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS) (pp. 1-7). IEEE. https://doi.org/10.1109/NETAPPS63333.2024.10823624

[32] Zhao, F., Zhang, C., & Geng, B. (2024). Deep multimodal data fusion. ACM computing surveys, 56(9), 1-36. https://doi.org/10.1145/3649447

[33] Luo, Y., Liu, W., Li, H., Lu, Y., & Lu, B. L. (2024). A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. Journal of Neural Engineering, 21(4), 046004. https://doi.org/10.1088/1741-2552/ad546d

[34] Alikhani, H., Kanduri, A., Liljeberg, P., Rahmani, A. M., & Dutt, N. (2023). Dynafuse: Dynamic fusion for resource efficient multimodal machine learning inference. IEEE Embedded Systems Letters, 15(4), 222-225. https://doi.org/10.1109/LES.2023.3298738

[35] Natan, O., & Miura, J. (2022). Towards compact autonomous driving perception with balanced learning and multi-sensor fusion. IEEE Transactions on Intelligent Transportation Systems, 23(9), 16249-16266. https://doi.org/10.1109/TITS.2022.3149370