

Article

Not peer-reviewed version

Uncertainty-Aware Evidential Fusion for Multi-Modal Object Detection in Autonomous Driving

[Qihang Yang](#), [Yang Zhao](#)^{*}, Hong Cheng

Posted Date: 7 January 2026

doi: 10.20944/preprints202601.0488.v1

Keywords: multi-modal fusion; 3D object detection; uncertainty estimation; autonomous driving; late fusion; deep learning



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Uncertainty-Aware Evidential Fusion for Multi-Modal Object Detection in Autonomous Driving

Qihang Yang¹, Yang Zhao^{2,*} and Hong Cheng²

¹ Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China

² School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

* Correspondence: yzhao@uestc.edu.cn

Highlights

What are the main findings?

- The MMLF late-fusion framework achieves significant, simultaneous performance gains for multi-class (Car, Pedestrian, Cyclist) object detection.
- Integrating uncertainty estimation directly into the fusion process substantially reduces predictive uncertainty, providing a clear, quantifiable measure of detection reliability.

What are the implications of the main finding?

- This work provides a practical and flexible fusion paradigm that enhances perception system performance without requiring architectural changes to existing, well-optimized 2D and 3D detectors.
- The model's built-in uncertainty quantification makes multi-modal object detection more transparent and trustworthy, enhancing safety for autonomous driving applications.

Abstract

The advancement of autonomous driving technologies necessitates the development of sophisticated object detection systems capable of integrating heterogeneous sensor data to overcome the inherent limitations of unimodal approaches. While multi-modal fusion strategies offer promising solutions, they confront significant challenges including data alignment complexities in early fusion and computational burdens coupled with overfitting risks in deep fusion methodologies. We proposed a Multi-modal Multi-class Late Fusion (MMLF) framework that operates at the decision level. This late-fusion strategy preserves the architectural integrity of individual detectors and facilitates the flexible integration of diverse modalities. A key innovation of our approach is the incorporation of an evidence-theoretic uncertainty quantification mechanism, built upon Dempster-Shafer theory, which provides a mathematically grounded measure of confidence and achieves substantial uncertainty reduction of approximately 77% for cars, 76% for pedestrians, and 67% for cyclists, thereby significantly improving both the reliability and interpretability of object detection outcome. Comprehensive experimental evaluation on the KITTI benchmark dataset demonstrates that our method achieves substantial performance improvements across multiple metrics, including 2D detection, 3D localization, and bird's-eye view tasks. The framework reduces uncertainty estimates across different object categories. This work provides a versatile and scalable solution for multi-modal object detection that effectively addresses critical challenges in autonomous driving applications.

Keywords: multi-modal fusion; 3D object detection; uncertainty estimation; autonomous driving; late fusion; deep learning

1. Introduction

Autonomous driving systems represent one of the most transformative technological developments of the past decade, with object detection serving as a fundamental component for environmental

perception and navigation decision-making [1]. The operational demands of these systems require exceptionally robust and accurate detection capabilities across diverse and dynamic real-world scenarios. Traditional unimodal approaches, which predominantly rely on either visual data from cameras or point cloud information from LiDAR sensors, frequently encounter limitations due to their inherent sensor-specific constraints [2]. Cameras provide rich texture and color information but suffer from sensitivity to lighting conditions and limited depth perception, while LiDAR offers precise spatial measurements but lacks semantic context and struggles with adverse weather conditions [3].

The integration of complementary sensor modalities through multi-modal fusion has emerged as a promising paradigm to overcome these individual limitations. Multi-modal approaches leverage the synergistic advantages of heterogeneous sensors to achieve enhanced detection accuracy, improved robustness to environmental variations, and increased operational reliability [4]. However, the implementation of effective fusion strategies presents considerable challenges. Early fusion techniques, which combine raw sensor data at the input level, require precise spatial and temporal alignment while being vulnerable to sensor-specific noise characteristics [5]. Deep fusion methods, operating at the feature level within neural network architectures, often introduce substantial computational overhead and face challenges related to overfitting and model complexity [6].

Beyond the fundamental challenges of multi-modal integration, the critical nature of autonomous driving applications demands rigorous uncertainty quantification to ensure operational safety. Conventional detection systems typically provide point estimates without conveying the reliability or confidence of these predictions [7]. This limitation becomes particularly concerning in safety-critical scenarios where understanding prediction uncertainty is essential for risk assessment and fallback strategy implementation [8]. Recent advancements in uncertainty-aware deep learning have highlighted the importance of quantifying both epistemic (model) and aleatoric (data) uncertainty to enhance the trustworthiness of automated systems [9].

The field has evolved beyond merely establishing the importance of uncertainty quantification. Current research is actively pursuing more sophisticated, efficient, and deeply integrated approaches. For instance, novel frameworks for trajectory prediction demonstrate that both aleatoric and epistemic uncertainties can be efficiently quantified through a single forward pass, significantly improving computational efficiency for resource-constrained platforms [10]. Furthermore, advances in cross-modal representation learning, such as hierarchical and monotonicity-aware alignment, underscore the ongoing effort to achieve more robust and semantically consistent fusion [11]. These works exemplify the field's progression toward methods that are not only accurate but also computationally practical and semantically grounded.

Despite these advancements, a distinct gap persists. Many existing multi-modal fusion frameworks still treat uncertainty estimation as a separate or post-hoc module, rather than an integral component that actively guides the adaptive fusion process itself. There is a pronounced need for a unified framework that tightly couples dynamic, context-aware sensor fusion with real-time, interpretable uncertainty quantification—without incurring prohibitive computational cost. This work aims to address this gap by proposing a Multi-modal Multi-class Late Fusion (MMLF) framework that integrates evidence-theoretic uncertainty quantification within a flexible fusion architecture. Our approach makes several significant contributions:

- The proposal of a flexible late fusion architecture designed for multi-class detection scenarios. The proposed MMLF framework integrates various 2D and 3D detectors in a plug-and-play manner, preserving their original structures while enabling effective cross-modal fusion.
- The systematic integration of evidential deep learning for uncertainty-aware fusion. We leverage Dempster-Shafer theory and subjective logic to explicitly model and quantify classification uncertainty, moving beyond traditional softmax outputs to provide trustworthy confidence scores for autonomous driving.

- Extensive validation on the KITTI benchmark. We demonstrate that our framework achieves substantial performance gains in 2D, 3D, and BEV detection tasks while simultaneously reducing prediction uncertainty across all object categories.

The remainder of this paper is organized as follows: Section 2 provides a review of related work in multi-modal fusion and uncertainty estimation. Section 3 details the theoretical foundation and architectural design of our proposed framework. Section 4 presents experimental setup, results, and comparative analysis. Section 5 discusses implications, limitations, and future research directions, followed by concluding remarks in Section 6.

2. Related Work

2.1. Multi-Modal Fusion Strategies

The evolution of multi-modal fusion methodologies has progressed through several distinct paradigms, each addressing specific challenges in sensor data integration. The fundamental objective of multi-modal fusion is to combine information from diverse sensors to achieve performance superior to any single modality alone [4]. This integration can occur at various levels of abstraction, leading to the categorization of fusion approaches into early, deep, and late fusion strategies.

2.1.1. Early Fusion

Early fusion approaches, also referred to as data-level fusion, operate directly on raw sensor measurements before feature extraction. These methods aim to create a unified representation of the input data that captures complementary information from multiple sources. Vora et al. [5] proposed PointPainting, which projects LiDAR points onto semantic segmentation maps derived from images, augmenting each point with corresponding class probabilities. This approach demonstrated significant improvements in 3D object detection by leveraging the semantic richness of images and the spatial precision of LiDAR. Similarly, Wen et al. [12] introduced a point feature fusion module that extracts features from RGB images and fuses them with corresponding point cloud data without relying on complex backbone networks.

Despite their potential benefits, early fusion methods face substantial challenges. They require meticulous calibration and synchronization between sensors to ensure accurate spatial alignment. The fusion of raw data also makes these approaches susceptible to propagating sensor-specific noise and artifacts through the processing pipeline. Furthermore, early fusion architectures often lack flexibility, as they are typically designed for specific sensor configurations and may not generalize well to new modalities or changing operational conditions. However, late fusion mitigates these issues by processing sensor data independently and integrating information at the decision level, thereby eliminating the need for precise spatiotemporal alignment, preventing cross-sensor noise propagation, and improving adaptability to diverse modalities and dynamic conditions.

2.1.2. Deep Fusion

Deep fusion techniques emerged to address limitations of early fusion by operating at the intermediate feature levels within neural network architectures. These methods learn to combine abstract representations extracted from different modalities, potentially capturing more complex interactions and dependencies. Among recent advancements, FusionFormer [13] represents a notable example in this category, employing transformer architectures to fuse features from diverse views for accurate 3D pose estimation. The attention mechanisms in transformers enable the model to dynamically weight the importance of different features and modalities based on the context. Similarly, Zhao et al. [14] proposed SimpleBEV, a LiDAR-camera fusion framework that enhances the individual camera and LiDAR encoders. Their approach incorporates depth rectification using LiDAR cues and an auxiliary camera-only detection branch during training.

However, deep fusion approaches introduce their own set of challenges. The integration of features from multiple modalities often requires significant architectural modifications and additional

parameters, leading to increased computational complexity and memory requirements. This complexity can result in overfitting, particularly when training data is limited. Additionally, the intertwined nature of feature extraction and fusion makes these models less interpretable and more difficult to optimize compared to simpler fusion strategies.

2.1.3. Late Fusion

Late fusion strategies, operating at the decision level, have gained prominence due to their flexibility and robustness to modality-specific characteristics. These approaches process each modality independently and combine the final predictions or decisions, offering several advantages for practical applications. Figuerêdo et al. [15] demonstrated the effectiveness of late fusion in combining early and late fusion approaches for depression detection in social media, highlighting the adaptability of decision-level fusion. Pang et al. [16] introduced CLOCs, which combines 2D and 3D detection results through a novel fusion network that operates on the geometric relationships between candidate detections.

Recent work by Li et al. [17] proposed the Fully Sparse Fusion framework, which aligns 2D and 3D instance segmentation results to generate multi-modal instances for bounding box prediction. This approach demonstrates the ongoing innovation in late fusion methodologies, particularly in addressing the challenges of heterogeneous data representation and alignment. Late fusion offers practical benefits including modularity, as individual modality processors can be updated or replaced independently, and robustness to missing or unreliable modalities. However, traditional late fusion approaches often struggle to capture the rich interactions between modalities that occur at earlier processing stages.

2.2. Uncertainty Estimation

The integration of uncertainty estimation within deep learning frameworks has become increasingly important for safety-critical applications such as autonomous driving. Uncertainty quantification provides crucial information about the reliability of predictions, enabling more informed decision-making and risk assessment.

2.2.1. Bayesian Approaches

Bayesian methods provide a principled framework for uncertainty quantification by treating model parameters as probability distributions rather than fixed values. Blundell et al. [18] introduced Bayesian neural networks with variational inference, enabling practical uncertainty estimation in deep learning models. This approach captures both epistemic uncertainty (related to model parameters) and aleatoric uncertainty (related to data noise), providing a comprehensive view of prediction reliability. Kucukelbir et al. [19] advanced this field with automatic differentiation variational inference, making Bayesian methods more accessible and scalable to complex deep learning architectures. In related domains such as HDR imaging, similar uncertainty-aware frameworks have been developed to model one-to-many mappings and enhance high-frequency details [20], further demonstrating the broad applicability of probabilistic approaches in vision tasks.

2.2.2. Monte-Carlo Dropout (MC-Dropout)

Monte-Carlo Dropout, introduced by Gal and Ghahramani [21], provides a Bayesian approximation framework for posterior inference by treating dropout as a probabilistic sampling technique. This method enables effective uncertainty quantification in neural networks and has been widely adopted across various domains, including computer vision and natural language processing. For instance, Zhao et al. [22] applied MC-Dropout to estimate predictive uncertainty in semantic segmentation tasks, demonstrating its practical utility in vision-related applications. Recent advances continue to refine such uncertainty estimation techniques for safety-critical systems. Notably, Bethell et al. [23] introduced MC-CP, a hybrid method that combines an adaptive Monte Carlo dropout with conformal prediction. MC-CP addresses key limitations of prior uncertainty estimation methods—such as

high computational cost or overly conservative prediction sets—by dynamically adjusting dropout at runtime to save resources while producing robust, statistically rigorous uncertainty intervals.

2.2.3. Ensemble Methods

Deep ensemble approaches have demonstrated remarkable effectiveness in uncertainty estimation while maintaining computational practicality. Lakshminarayanan et al. [24] showed that simple ensembles of neural networks could provide well-calibrated uncertainty estimates without the computational overhead of full Bayesian inference. Molchanov et al. [25] further enhanced this approach by combining ensembles with variational dropout, creating a hybrid method that leverages the strengths of both techniques. Peng et al. [26] applied deep ensemble methods to YOLOv3 object detection, analyzing the relationship between uncertainty and confidence measures in detection outputs. Similar principles have been successfully applied in remote sensing; for example, Sharma and Saharia [27] utilized a novel deep ensemble strategy for SAR-based flood mapping, generating explicit uncertainty layers to assess prediction reliability in critical real-world scenarios.

2.2.4. Evidential Deep Learning

Recent advances in evidence-based deep learning have provided novel approaches for uncertainty quantification using Dempster-Shafer theory. Sensoy et al. [9] introduced evidential deep learning, which models uncertainty by treating network outputs as evidence masses for Dirichlet distributions. This approach provides a theoretically grounded framework for uncertainty quantification that goes beyond traditional softmax outputs. Zou et al. [28] developed EvidenceCap for medical image segmentation, demonstrating the practical utility of evidence-based uncertainty estimation in critical applications. Han et al. [29] proposed trusted multi-view classification, which dynamically integrates multiple views at the evidence level using Dirichlet distributions and Dempster-Shafer theory. The key distinction between evidential fusion and a simple combination of Dirichlet parameters is that the latter often ignores the inherent conflicts between different evidence sources and lacks a principled mechanism to measure the resultant uncertainty. Simple operations, such as averaging the parameters, do not account for or resolve contradictions in the evidence. In contrast, evidential fusion strictly follows Dempster's rule of combination, which explicitly identifies, quantifies, and redistributes the probability mass associated with conflicting evidence. This process not only integrates information but also yields a rigorous, well-calibrated measure of total uncertainty that incorporates both the strength of individual sources and the degree of disagreement among them.

Compared to Bayesian, MC-Dropout, and ensemble methods that estimate uncertainty through parameter distributions, multiple stochastic passes, or output variations, evidential theory explicitly models distributional uncertainty from a single deterministic output. This intrinsic uncertainty representation can be obtained without adding computational resources or modifying the model architecture, allowing it to seamlessly adapt to late fusion paradigms. This characteristic is particularly advantageous for late-fusion decision integration, enabling the coherent combination of evidence from different sources or views.

3. Methodology

This section presents our trusted multi-modal fusion framework for object detection. The overall architecture of the proposed system is illustrated in Figure 1 (a).

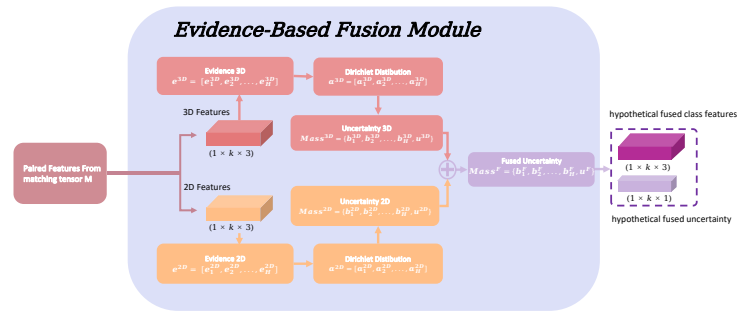
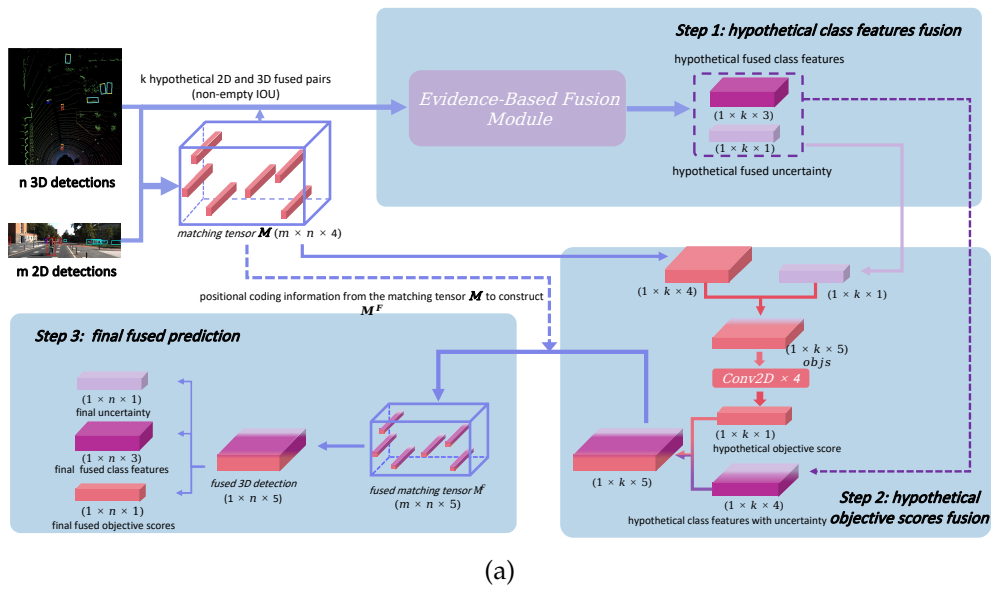


Figure 1. System architecture. (a) Flowchart of the proposed method: in step 1, each of the m 3D candidates is computed for IOU with each of the n 2D candidates to have k hypothetical fused pairs, and fused class features with uncertainty are obtained based on these pairs through Evidence-Based Fusion Module (detailed in (b)). In step 2, the hypothetical objective scores are computed by a 2D CNN and then concatenated with the fused class features with uncertainty which ultimately is used in step 3 to build fused matching tensor \mathbf{M}^F to get the final fused prediction. (b) Detailed structure of the Evidence-Based Fusion Module.

3.1. Multi-Modal Multi-Class Late Fusion (MMLF) Architecture

3.1.1. Hypothetical Class Features Fusion

The framework employed YOLOv8 [30] as the 2D detector and Complex-YOLO [31] as the 3D detector, both capable of multi-class detection required for effective fusion.

Let m and n represent the number of 3D and 2D detection candidates, respectively. We project the m 3D candidates onto the image plane and compute their Intersection-over-Union (IOU) with the n 2D candidates. This yields a matching tensor \mathbf{M} of dimensions $(m \times n \times 4)$. Each element in \mathbf{M} is defined as:

$$\mathbf{M}_{i,j} = \{ IOU_{i,j}^{3D-2D}, \mathbf{obj}s_i^{3D}, \mathbf{obj}s_j^{2D}, \mathbf{dis}_i^{3D} \} \quad (1)$$

Here, $IOU_{i,j}^{3D-2D}$ quantifies the spatial overlap between the i -th projected 3D candidate and j -th 2D candidate. $\mathbf{obj}s_i^{3D}$ and $\mathbf{obj}s_j^{2D}$ denote the objective scores of the respective candidates, while \mathbf{dis}_i^{3D} represents the normalized distance between the i -th 3D bounding box and the LiDAR sensor in the xy -plane.

From this process, k hypothetical 3D-2D pairs with non-zero IOU values are identified for fusion. The class features from these candidate pairs are treated as evidence [32], with 3D and 2D evidence vectors defined as $\mathbf{e}^{3D} = [e_1^{3D}, e_2^{3D}, \dots, e_H^{3D}]$ and $\mathbf{e}^{2D} = [e_1^{2D}, e_2^{2D}, \dots, e_H^{2D}]$, where H denotes the number of classes (e.g., $H = 3$ for KITTI Dataset).

Following subjective logic principles [33], these evidence vectors are transformed into Dirichlet distribution parameters $\boldsymbol{\alpha}^{3D} = [\alpha_1^{3D}, \alpha_2^{3D}, \dots, \alpha_H^{3D}]$ and $\boldsymbol{\alpha}^{2D} = [\alpha_1^{2D}, \alpha_2^{2D}, \dots, \alpha_H^{2D}]$. The corresponding belief masses $\mathbf{b}^{3D} = [b_1^{3D}, b_2^{3D}, \dots, b_H^{3D}]$, $\mathbf{b}^{2D} = [b_1^{2D}, b_2^{2D}, \dots, b_H^{2D}]$ and uncertainties u^{3D} , u^{2D} are computed using:

$$b_h = \frac{e_h}{S} = \frac{\alpha_h - 1}{S} \quad (\text{for } h = 1, 2, \dots, H) \quad (2)$$

$$u = \frac{H}{S} \quad (3)$$

where $S = \sum_i^H \alpha_i$ represents the Dirichlet strength.

The fused mass $\mathbf{Mass}^F = \{b_1^F, b_2^F, \dots, b_H^F, u^F\}$ is obtained through Dempster's combination rule [34,35], with fused belief masses and uncertainty calculated as:

$$b_h^F = \frac{1}{1-C} \left(b_h^{3D} b_h^{2D} + b_h^{3D} u^{2D} + b_h^{2D} u^{3D} \right) \quad (4)$$

$$u^F = \frac{1}{1-C} u^{3D} u^{2D} \quad (5)$$

where $C = \sum_{i \neq j} b_i^{3D} b_j^{2D}$ quantifies the conflict between 2D and 3D class detections.

3.1.2. Hypothetical Objective Scores Fusion

For each of the k hypothetical pairs, the fused uncertainty u^F - representing class conflict between modalities - is incorporated into the matching tensor \mathbf{M} to construct a $1 \times k \times 5$ objective score tensor:

$$\mathbf{objs} = \left\{ IOU_{i,j}^{3D-2D}, \mathbf{objs}_i^{3D}, \mathbf{objs}_j^{2D}, \mathbf{dis}_i^{3D}, \mathbf{u}_{i,j}^F \right\} \quad (6)$$

A series of 2D convolutional layers with (1,1) kernel size and unit stride progressively transform the third channel of \mathbf{objs} from $1 \times k \times 5$ to $1 \times k \times 18$, then to $1 \times k \times 36$, and finally to $1 \times k \times 1$, yielding the hypothetical objectness score vector. This layered expansion compresses basic features into increasingly abstract representations, enhancing the model's ability to model nonlinear interactions relevant to objectness. The choice of (1 × 1) convolutions—rather than fully-connected layers—reduces parameter counts and improves computational efficiency, while also aligning with typical image feature processing pipelines for easier integration and extension.

3.1.3. Final Fused Prediction

The hypothetical objective score vector ($1 \times k \times 1$) is concatenated with hypothetical class features and uncertainty information ($1 \times k \times 4$) to form the fused matching tensor \mathbf{M}^F . The entry with the highest objective score in the second dimension is selected using the Max function, providing the final fused 3D detection comprising objective scores, class features, and uncertainty. Application of Non-Maximum Suppression (NMS) and filtering of high-uncertainty detections yields more confident final predictions.

3.1.4. Illustrative Example

Here is a real and specific numerical example from the fusion process to illustrate the method, which helps readers better understand our approach.

Step 1: Pre-Fusion Mass Calculation

$$\begin{aligned}\mathbf{e}^{3D} &= [e_1^{3D}, e_2^{3D}, e_3^{3D}] = [1.3533 \times 10^1, 4.4121 \times 10^{-3}, 3.3779 \times 10^{-3}], \\ \mathbf{e}^{2D} &= [e_1^{2D}, e_2^{2D}, e_3^{2D}] = [2.1956 \times 10^1, 8.2941 \times 10^{-3}, 8.5092 \times 10^{-4}]\end{aligned}$$

Here, \mathbf{e}^{3D} and \mathbf{e}^{2D} represent the initial evidence in 3D and 2D spaces, by adding one to each element of the initial vectors, we obtain the Dirichlet distribution parameters:

$$\begin{aligned}\boldsymbol{\alpha}^{3D} &= \mathbf{e}^{3D} + \mathbf{1} = [1.4533 \times 10^1, 1.0044, 1.0034], \\ \boldsymbol{\alpha}^{2D} &= \mathbf{e}^{2D} + \mathbf{1} = [2.2956 \times 10^1, 1.0083, 1.0009]\end{aligned}$$

Then, we introduce the Dirichlet strength for both cases:

$$S^{3D} = \sum_i^3 \alpha_i^{3D} = 1.6541 \times 10^1, \quad S^{2D} = \sum_i^3 \alpha_i^{2D} = 2.4965 \times 10^1$$

Based on Dirichlet strength, we calculate the belief masses:

$$\begin{aligned}\mathbf{b}^{3D} &= \left[\frac{e_1^{3D}}{S^{3D}}, \frac{e_2^{3D}}{S^{3D}}, \frac{e_3^{3D}}{S^{3D}} \right] = [0.81816, 2.6673 \times 10^{-4}, 2.0421 \times 10^{-4}], \\ \mathbf{b}^{2D} &= \left[\frac{e_1^{2D}}{S^{2D}}, \frac{e_2^{2D}}{S^{2D}}, \frac{e_3^{2D}}{S^{2D}} \right] = [0.87947, 3.3223 \times 10^{-4}, 3.4085 \times 10^{-5}]\end{aligned}$$

The uncertainties u^{3D} and u^{2D} are computed as $3/S^{3D}$ and $3/S^{2D}$, respectively:

$$u^{3D} = \frac{3}{S^{3D}} = 0.18137, \quad u^{2D} = \frac{3}{S^{2D}} = 0.12017.$$

Finally, the Masses for fuse are combined as:

$$\begin{aligned}\text{Mass}^{3D} &= [b_0^{3D}, b_1^{3D}, b_2^{3D}, u^{3D}] = [0.8182, 2.7 \times 10^{-4}, 2.0 \times 10^{-4}, 0.1814], \\ \text{Mass}^{2D} &= [b_0^{2D}, b_1^{2D}, b_2^{2D}, u^{2D}] = [0.8795, 3.3 \times 10^{-4}, 3.4 \times 10^{-5}, 0.1202].\end{aligned}$$

Step 2: Fusion Calculation

Based on the belief masses calculated in Step 1, we first compute the conflict factor C:

$$C = \sum_{i \neq j} b_i^{3D} b_j^{2D} = 7.14 \times 10^{-4} (1 - C = 0.9993)$$

Then, applying Dempster's combination rule, we compute the fused belief masses and uncertainty:

$$\begin{aligned}b_0^F &= \frac{1}{1-C} (b_0^{3D} b_0^{2D} + b_0^{3D} u^{2D} + b_0^{2D} u^{3D}) = 0.9781, \\ b_1^F &= \frac{1}{1-C} (b_1^{3D} b_1^{2D} + b_1^{3D} u^{2D} + b_1^{2D} u^{3D}) = 9.87 \times 10^{-5}, \\ b_2^F &= \frac{1}{1-C} (b_2^{3D} b_2^{2D} + b_2^{3D} u^{2D} + b_2^{2D} u^{3D}) = 3.08 \times 10^{-5}, \\ u^F &= \frac{1}{1-C} u^{3D} u^{2D} = 0.0218\end{aligned}$$

Finally, the fused mass is:

$$\mathbf{M}^F = [b_0^F, b_1^F, b_2^F, u^F] = [0.9781, 9.87 \times 10^{-5}, 3.08 \times 10^{-5}, 0.0218].$$

After fusion, the fused mass supporting classification as Class 0 becomes more robust, accompanied by a reduced level of uncertainty.

3.2. Non-Matching Scenarios

To mitigate the over-reliance on 2D evidence and prevent the outright dismissal of valuable but unpaired 3D detections, a virtual pair mechanism is introduced for projected 3D candidates with no intersecting 2D candidate. In these non-matching scenarios, the 2D component is simulated by setting its confidence score to a low value (e.g., -10) and its evidence vector to zero. During fusion, this configuration ensures the combined belief masses and uncertainty are dominantly governed by the original, confident 3D modality, thereby preserving its informational value while appropriately reflecting the lack of 2D corroboration through a heightened uncertainty score. This approach aims to balance multimodal integration against the risk of being constrained by the limitations of a single sensor stream.

3.3. Uncertainty Estimation

Our approach initially performs hypothetical fusion based on IOU information from matching tensor \mathbf{M} , deriving hypothetical fused uncertainty that captures inter-modal conflicts. This uncertainty is incorporated into the objective score tensor during hypothetical objective scores fusion to reduce selection of illogical fusions. In the final prediction stage, combining with fused objective scores enhances representational capacity and interpretability, enabling discrimination of the network's confidence in both object localization and categorization. Subsequent filtering of high-uncertainty detections eliminates unreasonable results for more reliable decision-making.

3.4. Optimization and Training Strategy

We employ a multi-objective optimization approach that jointly minimizes detection errors and uncertainty estimates. The loss function combines several components to address different aspects of the fusion process:

The objective loss \mathcal{L}_{obj} employs the binary cross-entropy loss \mathcal{L}_{BCE} to ensure accurate detection and localization, which is computed using the objective mask and no-objective mask as follows:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\text{BCE}}(\text{obj_mask}) + \mathcal{L}_{\text{BCE}}(\text{no_obj_mask}) \quad (7)$$

The classification loss utilizes the sample-specific loss \mathcal{L}_{ssl} from [29].

$$\mathcal{L}_{\text{cls}} = \sum_{i=1}^N [\mathcal{L}_{\text{ssl}}(\alpha_i^F) + \mathcal{L}_{\text{ssl}}(\alpha_i^{3D}) + \mathcal{L}_{\text{ssl}}(\alpha_i^{2D})] \quad (8)$$

For non-matching scenarios where no corresponding 2D detection exists, we preserve the original 3D detection and apply a modified loss:

$$\mathcal{L}_{\text{non-match}} = \sum_{i=1}^N \mathcal{L}_{\text{ssl}}(\alpha_i^{3D})_{\text{non-match}} \quad (9)$$

The total loss is the weighted combination:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{obj}} + \lambda_2 \mathcal{L}_{\text{cls}} + \lambda_3 \mathcal{L}_{\text{non-match}} \quad (10)$$

where the coefficients are set to $\lambda_1 = 0.8$, $\lambda_2 = 1$, and $\lambda_3 = 1$ to balance the contribution of each loss component, a configuration that has been found to be effective in practice.

We utilize the Adam optimizer with learning rate 0.003 for evidence fusion components and Stochastic Gradient Descent with the same learning rate for objective score fusion. The training process emphasizes balanced learning across all components to ensure stable convergence and optimal performance.

4. Experimental Evaluation

4.1. Experimental Setup

All experiments were conducted on a workstation equipped with an Intel Core i5-12600KF CPU and an NVIDIA GeForce RTX 4080 GPU with 16GB of VRAM. The system had 32GB of RAM. During inference, our framework processed samples at an average speed of 0.512 seconds per sample. The training phase averaged 2.134 seconds per sample, and the model was trained for 5 epochs. The peak GPU memory usage during training was approximately 3733 MB.

We conducted comprehensive experiments on the KITTI dataset [36], which represents a standard benchmark for autonomous driving perception tasks. The dataset provides synchronized data from LiDAR and camera sensors, with detailed annotations for 2D and 3D object detection tasks. In detail, we split the 7481 training samples into 6000 samples for training and 1481 for validation. Additionally, we evaluated our method on the official test set comprising 7518 samples, with results submitted to the KITTI evaluation server.

For 3D object detection, we employed Complex-YOLO v3 and v4 as baseline detector. These models process LiDAR point cloud data to generate 3D bounding boxes. For 2D detection, we utilized YOLOv8 [30]. This detector provides high-quality 2D bounding boxes and class predictions from image data. No additional data processing steps were introduced. That is, for point cloud data, we directly followed the processing pipeline of Complex-YOLO, and for image data, we adhered to the standard procedure of YOLOv8.

The evaluation metrics included Average Precision (AP) with the unit being percentage (%) for 2D detection, 2D orientation estimation, 3D detection, and bird's-eye view (BEV) tasks. We reported results for three difficulty levels (Easy, Moderate, Hard) across three object categories (Car, Pedestrian, Cyclist), following the standard KITTI evaluation protocol. Uncertainty metrics were evaluated using the uncertainty scores derived from our evidence-theoretic framework.

4.2. Comparative Analysis

A comprehensive performance comparison between the baseline 3D-only detection (Ori) and the proposed multimodal fusion method (Fuse) is detailed in Table 1 (validation set) and Table 2 (test set). The analysis evaluates four key metrics—2D Detection, 2D Orientation, Bird's-Eye-View (BEV) detection, and 3D detection—and categorizes the findings into significant improvements, minor improvements or parity, and notable regressions.

Table 1. KITTI Validation Dataset Performance Comparison.

Methods	Car			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
eval 2D Detection AP (%)									
Ori-v3	56.85	50.94	53.23	24.51	26.57	24.50	54.38	43.52	44.14
Fuse-v3	65.31	57.48	58.89	38.43	37.83	37.77	66.40	54.54	54.35
Improvement	+8.46	+6.54	+5.66	+13.92	+11.26	+13.27	+12.02	+11.02	+10.21
Ori-v4	56.46	51.72	54.74	49.17	52.11	53.50	55.19	43.99	44.25
Fuse-v4	67.80	59.12	60.48	68.74	62.95	62.84	66.92	61.32	55.36
Improvement	+11.34	+7.40	+5.74	+19.57	+10.84	+9.34	+11.73	+17.33	+11.11

Table 1. Cont.

Methods	Car			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
eval 2D Orientation AP (%)									
Ori-v3	56.84	50.87	53.16	24.09	26.12	24.19	53.66	43.30	43.93
Fuse-v3	65.30	57.46	58.84	35.31	34.75	34.94	66.26	54.43	54.24
Improvement	+8.46	+6.59	+5.68	+11.22	+8.63	+10.75	+12.60	+11.13	+10.31
Ori-v4	56.45	51.68	54.70	47.71	50.57	51.83	55.10	43.94	44.19
Fuse-v4	67.79	59.09	60.42	67.12	61.26	60.35	66.84	61.25	55.30
Improvement	+11.34	+7.41	+5.72	+19.41	+10.69	+8.52	+11.74	+17.31	+11.11
BEV AP (%)									
Ori-v3	89.68	89.02	89.12	51.96	54.25	55.84	63.94	71.47	71.95
Fuse-v3	89.92	89.30	89.44	64.67	65.35	66.34	69.29	77.29	77.22
Improvement	+0.24	+0.28	+0.32	+12.71	+11.10	+10.50	+5.35	+5.82	+5.27
Ori-v4	89.46	89.33	89.43	56.71	54.57	55.88	75.25	76.51	76.41
Fuse-v4	90.06	89.61	89.61	73.75	69.26	69.08	76.72	77.25	77.40
Improvement	+0.60	+0.28	+0.18	+17.04	+14.69	+13.20	+1.47	+0.74	+0.99
eval 3D detection AP (%)									
Ori-v3	29.50	27.74	29.70	30.34	28.68	30.61	19.95	19.87	20.82
Fuse-v3	38.43	34.62	35.91	64.67	65.35	66.34	69.29	77.29	77.22
Improvement	+8.93	+6.88	+6.21	+34.33	+36.67	+35.73	+49.34	+57.42	+56.40
Ori-v4	37.04	36.13	39.23	43.06	40.77	42.10	39.16	29.87	30.47
Fuse-v4	50.75	44.62	45.50	58.38	54.07	49.25	51.15	40.91	40.75
Improvement	+13.71	+8.49	+6.27	+15.32	+13.30	+7.15	+11.99	+11.04	+10.28

Table 2. KITTI Test Dataset Performance Comparison.

Methods	Car			Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Test 2D Detection AP (%)									
Ori-v3	47.02	37.02	37.01	10.06	7.07	7.19	28.73	21.08	18.32
Fuse-v3	58.07	44.00	43.20	21.69	14.47	13.03	36.64	26.94	24.90
Improvement	+11.05	+6.98	+6.19	+11.63	+7.40	+5.84	+7.91	+5.86	+6.58
Ori-v4	44.72	38.36	37.88	27.17	19.97	18.74	21.32	15.01	13.50
Fuse-v4	61.58	45.67	44.47	39.56	28.77	26.77	36.21	26.69	23.90
Improvement	+16.86	+7.31	+6.59	+12.39	+8.80	+8.03	+14.89	+11.68	+10.40
Test 2D Orientation AP (%)									
Ori-v3	47.00	36.90	36.85	6.27	4.42	4.48	27.80	19.42	16.95
Fuse-v3	58.03	43.81	42.93	13.48	9.13	8.13	34.35	23.81	22.02
Improvement	+11.03	+6.91	+6.08	+7.21	+4.71	+3.65	+6.55	+4.39	+5.07
Ori-v4	44.60	38.12	37.60	21.33	15.37	14.27	34.59	25.47	22.80
Fuse-v4	61.32	45.41	44.17	30.96	22.18	20.38	38.03	28.53	25.12
Improvement	+16.72	+7.29	+6.57	+9.63	+6.81	+6.11	+3.44	+3.06	+2.32
Test BEV AP (%)									
Ori-v3	83.02	74.39	66.34	16.13	12.62	11.55	27.53	19.22	17.65
Fuse-v3	80.24	72.91	67.26	20.16	15.17	13.87	32.24	26.42	23.34
Improvement	-2.78	-1.48	+0.92	+4.03	+2.55	+2.32	+4.71	+7.20	+5.69
Ori-v4	81.88	75.38	69.42	25.00	18.42	17.20	39.47	30.80	27.62
Fuse-v4	84.23	77.99	71.74	33.46	25.96	23.15	41.28	34.98	29.68
Improvement	+2.35	+2.61	+2.32	+8.46	+7.54	+5.95	+1.81	+4.18	+2.06
Test 3D bounding boxes AP (%)									
Ori-v3	16.25	13.39	13.42	7.22	5.33	4.64	8.57	5.43	5.46
Fuse-v3	21.51	16.23	15.92	12.08	8.57	7.56	11.95	8.49	7.67
Improvement	+5.26	+2.84	+2.50	+4.86	+3.24	+2.92	+3.38	+3.06	+2.21
Ori-v4	20.41	18.26	18.08	13.77	10.23	9.24	13.85	9.10	8.53
Fuse-v4	31.32	23.82	22.94	21.32	15.01	13.50	17.90	12.16	10.55
Improvement	+10.91	+5.56	+4.86	+7.55	+4.78	+4.26	+4.05	+3.06	+2.02

4.2.1. Significant Improvements

The most substantial gains are observed in tasks where complementary 2D visual features provide critical information missing from LiDAR data alone.

2D Detection & Orientation Accuracy: The Fuse model achieves significant and consistent improvements in 2D performance across all object categories and difficulty levels. For example, on the test set, the 2D Detection AP for the Car class using Fuse-v4 increased by 16.86, 7.31, and 6.59 percentage points for Easy, Moderate, and Hard levels, respectively. Gains for the Pedestrian and Cyclist classes on the validation set are even more pronounced, frequently exceeding 10 percentage points. This demonstrates that integrating image features dramatically enhances object recognition and localization within the 2D plane.

3D Detection Accuracy: The advantage of sensor fusion is most transformative for 3D object detection, particularly for smaller and more challenging classes. On the validation set, 3D AP for Pedestrian and Cyclist using Fuse-v3 surged by over 34 and 49 percentage points, respectively. Although the v4 baseline is stronger, Fuse-v4 still delivers major improvements, such as a 15.32 percentage point gain for Pedestrian (Easy). These results underscore that 2D semantic and contextual information is vital for accurately estimating the 3D pose of objects that are poorly represented in sparse point clouds.

4.2.2. Minor Improvements or Parity

Metrics that rely heavily on precise 3D geometry show more modest gains, as the baseline LiDAR-based performance is already strong.

Bird's-Eye-View (BEV) Detection: For the Car class, which exhibits very high baseline BEV AP (often >89%), the fusion approach yields only minor improvements, typically less than 1 percentage point. This indicates that for large objects with abundant LiDAR points, BEV detection is nearly saturated. Improvements for Pedestrian and Cyclist in BEV are consistent but generally smaller in magnitude than their corresponding 3D AP gains, suggesting that fusion helps but does not fully overcome the inherent challenge of precise BEV localization for small objects.

4.2.3. Minor Reductions

BEV AP for Car on Test Set (Fuse-v3): In contrast to the overall positive trend, Fuse-v3 exhibited a decrease in BEV AP for the Car class at Easy and Moderate difficulty levels on the test set, dropping by 2.78 and 1.48 percentage points, respectively. This regression was corrected in the Fuse-v4 model, which shows positive gains. The primary cause lies with the original 3D model (v3), which suffered from overfitting and generated a slightly excessive number of false bounding boxes for cars, thereby leading to erroneous fusion. When the base model was upgraded to v4, the overall quality of the output bounding boxes improved significantly.

In conclusion, the multimodal fusion method delivers significant improvements in 2D and 3D detection, especially for smaller objects like pedestrians and cyclists, by effectively combining LiDAR and image data. Gains in BEV detection are more modest for larger objects, where LiDAR-only performance is already strong. An initial regression observed in early model version was addressed in the enhanced version, confirming the importance of a robust 3D detection backbone. Overall, the results affirm the potential of sensor fusion for accurate perception.

4.3. Uncertainty Analysis

The proposed framework achieves significant uncertainty reduction while maintaining detection accuracy, as summarized in Table 3. The evidence-theoretic uncertainty quantification provides mathematically grounded confidence measures that enhance the reliability of detection outcomes.

For Complex-YOLO v3, the fusion framework reduces average uncertainty scores from 0.11827 to 0.02692 for cars (77.3% reduction), from 0.24373 to 0.05354 for pedestrians (78.0% reduction), and from 0.23636 to 0.07354 for cyclists (68.9% reduction). Similar reductions are observed for the v4 variant, demonstrating the consistency of our uncertainty quantification approach.

Table 3. Average uncertainty score comparison.

Method \ Class	car	pedestrian	cyclist
Ori-v3	0.1182	0.2437	0.2363
Fuse-v3	0.0269	0.0535	0.0735
Ori-v4	0.1188	0.2503	0.2395
Fuse-v4	0.0273	0.0579	0.0792

The uncertainty reduction is particularly significant for challenging categories like pedestrians and cyclists, where single-modality detectors often exhibit high uncertainty due to limited sensor information. The multi-modal fusion effectively combines complementary information, leading to more confident and reliable predictions. This uncertainty reduction has important implications for safety-critical applications, where understanding prediction reliability is essential for risk assessment and decision-making.

4.4. Ablation Studies

To systematically verify the effectiveness of each component in our proposed fusion framework, we conducted a comprehensive ablation study on the KITTI dataset. The experimental results, detailed in Table 4, were obtained by incrementally integrating the hypothetical Class Fusion (CF) module and the hypothetical Objective Scores Fusion (OSF) module into the baseline Complex-YOLOv3 network. From the ablation study results, several key observations can be made regarding the performance improvements across different Complex-YOLO versions.

Table 4. Ablation study on different baseline models. ‘CF’ stands for hypothetical Class Fusion, and ‘OSF’ represents hypothetical objective scores fusion. A check mark(✓) indicates that the corresponding feature is used. mAP (mean Average Precision) is the primary evaluation metric for detection performance. Best results for each baseline version are in bold. Results with Δ are from Complex-YOLOv3, and with \square are from Complex-YOLOv4.

Baseline	CF	OSF	mAP (%)			
			2D Detection	2D Orientation	Bird’s Eye View	3D Detection
✓ Δ	-	-	44.24	44.07	79.02	27.48
✓ Δ	✓	-	48.07	47.62	79.11	34.02
✓ Δ	-	✓	51.54	51.01	82.06	37.01
✓ Δ	✓	✓	52.26	51.46	82.42	45.14
✓ \square	-	-	51.33	50.91	79.63	36.92
✓ \square	✓	-	54.59	54.11	80.16	43.21
✓ \square	-	✓	57.40	56.92	82.00	44.77
✓ \square	✓	✓	60.23	59.78	83.63	46.80

The incorporation of the CF module consistently enhanced detection performance in both versions. In Complex-YOLOv3, CF increased 2D detection mAP from 44.24% to 48.07% and boosted 3D detection mAP from 27.48% to 34.02%. Similarly, in Complex-YOLOv4, CF improved 2D detection from 51.33% to 54.59% and 3D detection from 36.92% to 43.21%, demonstrating the effectiveness of class fusion across different baseline architectures.

The OSF module alone provided even more substantial gains. For Complex-YOLOv3, OSF elevated 2D detection mAP to 51.54% and 3D detection to 37.01%, while in Complex-YOLOv4, it achieved 57.40% in 2D detection and 44.77% in 3D detection. This confirms that objective scores fusion significantly optimizes detection confidence across both versions.

The complete framework integrating both CF and OSF modules achieved the best performance in all evaluation metrics. For Complex-YOLOv3, the full model reached 52.26% in 2D detection, 51.46% in orientation estimation, 82.42% in bird’s-eye view, and 45.14% in 3D detection. Complex-YOLOv4

with both modules performed even better, achieving 60.23% in 2D detection, 59.78% in orientation, 83.63% in bird's-eye view, and 46.80% in 3D detection.

Notably, the performance gap between versions highlights the architectural improvements in Complex-YOLOv4, which consistently outperformed Complex-YOLOv3 across all configurations. The most significant absolute improvements were observed in 3D detection, where the full Complex-YOLOv4 model achieved a 17.66% gain over the baseline Complex-YOLOv3, and the full Complex-YOLOv4 showed a 9.88% improvement over its own baseline.

These results collectively validate that our fusion framework effectively leverages complementary information and can be successfully integrated with different baseline architectures, with the combined CF and OSF modules providing the most robust performance across all detection tasks.

4.5. Discussion

4.5.1. Model Significance and Multimodal Fusion

The model's significance in autonomous driving detection is multifaceted. Firstly, it significantly enhances detection and positioning accuracy by seamlessly fusing 2D and 3D information, thereby improving the overall performance of the sensing module. Notably, it excels in accurately predicting pedestrian directions, a critical capability for autonomous driving systems to anticipate pedestrian behavior and enhance safety in diverse traffic scenarios. Secondly, the model demonstrated strong generalization ability, showcasing robust performance not only on verification sets but also on test sets. This underscores its adaptability across various scenarios and conditions, providing a reliable solution for real-world deployment in autonomous driving systems. Moreover, the model seamlessly integrates information from a wider range of detection patterns, efficiently fusing object-level information from multiple modalities without altering its original structure. This adaptability enables the model to capture richer environmental representations in autonomous driving scenarios. Additionally, the model incorporates uncertainty estimation, enabling the evaluation of detection reliability based on uncertainty scores—a critical aspect for ensuring safe and efficient driving systems. This capability is further enhanced by the application of frameworks like Dempster-Shafer theory, whose key benefit is the ability to explicitly quantify cross-modal conflict, transforming sensor disagreement into actionable information. This enables adaptive fusion strategies: for example, the system can dynamically weigh LiDAR data more heavily for large objects like vehicles while relying more on image data for smaller objects like pedestrians, especially when conflict is detected. Crucially, these strategies also address scenarios where high-quality 2D support is unavailable. When fusion fails due to insufficient 2D evidence—leading to a high uncertainty score—yet the 3D modality reports a high objective score, the fusion rule retains the detection. In such cases, it appropriately reduces the final objective score while keeping the uncertainty flag elevated. This directs attention to potentially ambiguous detections, proactively prompting safer decision-making—such as more conservative driving maneuvers or a request for human oversight. By explicitly modeling and utilizing uncertainty in this principled manner, the model achieves more robust, context-aware decision-making, contributing directly to safer autonomous operation.

4.5.2. Visual Results and Uncertainty-Relevant Driving Safety

Furthermore, the effectiveness of our fusion framework is visually demonstrated in Figure 2, which presents comparative Bird's Eye View results on the KITTI validation set. The visualized comparisons reveal three key advantages of our approach: (a) In handling occluded objects, the baseline model (ori3d) exhibits high uncertainty and missed detections for partially occluded vehicles, while our fused approach (fuse3d) successfully detects these occluded targets with considerably lower uncertainty, showcasing the framework's robustness to occlusion challenges through effective multi-modal feature integration. (b) For small object detection, the baseline detector shows either complete miss-detections or high uncertainty values for distant and small-scale objects, whereas our MMLF framework demonstrates remarkable improvement in detecting these challenging small objects with confident low-uncertainty predictions, validating the effectiveness of our scale-aware fusion strategy. (c) In long-range scenarios, where traditional 3D detectors typically suffer from performance

degradation, our model maintains reliable detection capability. This improvement stems from a targeted fusion strategy: rather than performing independent detection, the 2D stream is tasked specifically with augmenting the confidence of the initial 3D hypotheses. By supplying rich semantic context where LiDAR data is sparse, it refines these predictions, thereby enabling the 3D detector to operate more reliably.

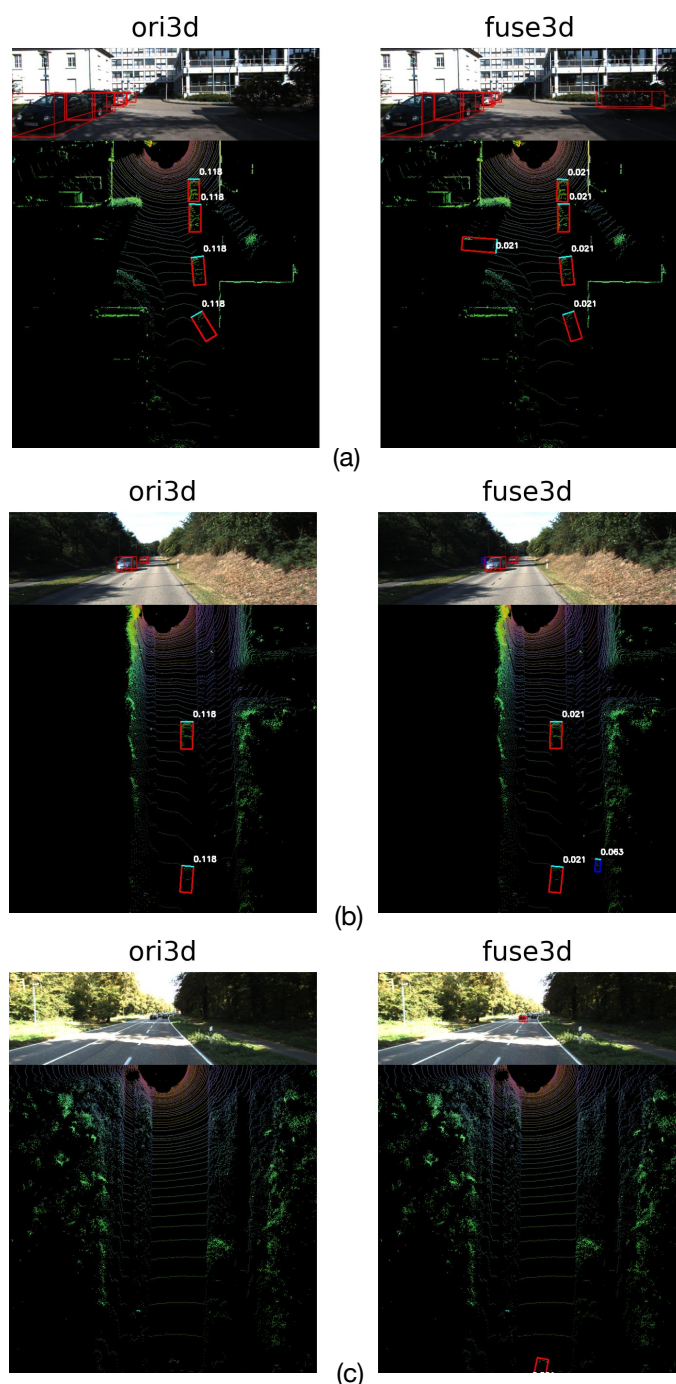


Figure 2. Visualized results of our MMLF on KITTI validation set. In the 3D BEV representation, the numbers adjacent to the bounding boxes indicate uncertainty. Our model has significantly reduced the original detection uncertainty and demonstrated an improvement in detection capability. (a) Demonstrates the enhanced detection capability of our fused model in the presence of occluded objects. (b) Illustrates the improved detection ability of our model for small objects. (c) Indicates that our model partially addresses the issue of weak long-range object detection inherent in 3D detectors.

Beyond improved detection performance, our framework quantifies prediction reliability through calibrated uncertainty scores. Low uncertainty signifies consistent and mutually reinforcing evidence across sensor modalities, supporting confident and reliable detection. Conversely, high uncertainty indicates sensor conflict or insufficient information, which can trigger critical safety protocols—such as requesting human intervention or switching to a more cautious driving mode. For instance, reducing pedestrian-related uncertainty from 0.24 to 0.05 means the system now aggregates strong, consistent evidence from multiple sensors confirming a pedestrian's presence. In practice, this high confidence enables timely and appropriate actions, such as smooth, early braking, while effectively avoiding two dangerous failure modes: missing a real pedestrian (a potential LiDAR-only blind spot) or braking for a non-existent one (a camera-only error common in poor lighting). Furthermore, the system can dynamically adjust its driving strategy based on the uncertainty and context. It will react most sensitively to uncertainty changes for a nearby pedestrian, potentially executing immediate braking, but respond more calmly to uncertainty variations for a distant vehicle, thereby ensuring both safety and driving efficiency. These visual results and uncertainty interpretations provide compelling evidence of our method's practical applicability in autonomous driving contexts, addressing critical perception and reliability challenges essential for real-world deployment.

5. Conclusion

This research has presented a comprehensive Multi-modal Multi-class Late Fusion framework with integrated uncertainty estimation for robust 3D object detection. Our approach addresses critical challenges in autonomous driving perception by providing a flexible fusion architecture that preserves detector integrity while enabling effective multi-modal integration. The incorporation of evidence-theoretic uncertainty quantification enhances the reliability and transparency of detection outcomes, providing crucial confidence measures for safety-critical applications.

Experimental results demonstrate that our framework achieves performance on standard benchmarks while significantly reducing uncertainty estimates across all object categories. The method's modular design facilitates integration with existing detection systems and provides a foundation for future advancements in multi-modal perception.

The evidence-theoretic foundation provides a mathematically rigorous framework for uncertainty-aware fusion, offering advantages over traditional probabilistic approaches. The practical implementation demonstrates real-time capability and compatibility with diverse detection architectures, enhancing its applicability in real-world autonomous driving systems.

6. Limitations and Future Work

Despite its compelling performance, our work has certain limitations. First, the fusion efficacy is contingent on the quality of the initial 2D and 3D proposals; poor performance from either upstream detector can propagate through our fusion network. Second, while the framework is designed for robustness, its performance under extreme sensor degradation (e.g., complete camera occlusion or LiDAR failure) has not been fully explored. Finally, the current method operates on a single frame, not leveraging temporal information. Our future work will focus on addressing these limitations by investigating robust fusion under sensor failure, incorporating temporal dynamics for more stable tracking, and extending the evidential framework to also quantify spatial localization uncertainty.

Author Contributions: Conceptualization, Q.Y. and Y.Z.; methodology, Q.Y.; software, Q.Y.; validation, Q.Y. and Y.Z.; formal analysis, Q.Y.; investigation, Q.Y.; resources, Y.Z. and H.C.; data curation, Q.Y.; writing—original draft preparation, Q.Y.; writing—review and editing, Y.Z. and H.C.; visualization, Q.Y.; supervision, Y.Z. and H.C.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2022YFB2503004), the Sichuan Regional Innovation Cooperation Project (NO. 2025YFHZ0291), and the Opening

Project of International Joint Research Center for Robotics and Intelligence System of Sichuan Province (Grant JQZN2023-005).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author. The KITTI dataset used in this study is publicly available at <http://www.cvlibs.net/datasets/kitti/>. Code is available at <https://github.com/Yang-Qihang11/MMLF>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liang, L.; Ma, H.; Zhao, L.; Xie, X.; Hua, C.; Zhang, M.; Zhang, Y. Vehicle detection algorithms for autonomous driving: A review. *Sensors* **2024**, *24*, 3088.
2. Xu, G.; Khan, A.S.; Moshayedi, A.J.; Zhang, X.; Shuxin, Y. The Object Detection, Perspective and Obstacles In Robotic: A Review. *EAI Endorsed Transactions on AI and Robotics* **2022**, *1*.
3. Wang, H.; Liu, J.; Dong, H.; Shao, Z. A Survey of the Multi-Sensor Fusion Object Detection Task in Autonomous Driving. *Sensors* **2025**, *25*, 2794.
4. Huang, K.; Shi, B.; Li, X.; Li, X.; Huang, S.; Li, Y. Multi-Modal Sensor Fusion for Auto Driving Perception: A Survey. *arXiv* **2022**. *arXiv preprint arXiv:2202.02703*.
5. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3d object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4604–4612.
6. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3d object detection. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer, 2020, pp. 35–52.
7. Kraus, F.; Dietmayer, K. Uncertainty estimation in one-stage object detection. In Proceedings of the 2019 IEEE intelligent transportation systems conference (ITSC). IEEE, 2019, pp. 53–60.
8. He, Y.; Zhu, C.; Wang, J.; Savvides, M.; Zhang, X. Bounding box regression with uncertainty for accurate object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2888–2897.
9. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* **2018**, *31*.
10. Suk, H.; Kim, S. Uncertainty-Aware Multimodal Trajectory Prediction via a Single Inference from a Single Model. *Sensors (Basel, Switzerland)* **2025**, *25*, 217.
11. Wu, R.; Chen, P.; Shen, F.; Zhao, S.; Hui, Q.; Gao, H.; Lu, T.; Liu, Z.; Zhao, F.; Wang, K.; et al. HiMo-CLIP: Modeling Semantic Hierarchy and Monotonicity in Vision-Language Alignment. *arXiv preprint arXiv:2511.06653* **2025**.
12. Wen, L.H.; Jo, K.H. Fast and accurate 3D object detection for lidar-camera-based autonomous vehicles using one shared voxel-based backbone. *IEEE access* **2021**, *9*, 22080–22089.
13. Cai, Y.; Zhang, W.; Wu, Y.; Jin, C. Fusionformer: A concise unified feature fusion transformer for 3d pose estimation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 900–908.
14. Zhao, Y.; Gong, Z.; Zheng, P.; Zhu, H.; Wu, S. Simplebev: Improved lidar-camera fusion architecture for 3d object detection. *arXiv preprint arXiv:2411.05292* **2024**.
15. Figuerêdo, J.S.L.; Maia, A.L.L.; Calumby, R.T. Early depression detection in social media based on deep learning and underlying emotions. *Online Social Networks and Media* **2022**, *31*, 100225.
16. Pang, S.; Morris, D.; Radha, H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10386–10393.
17. Li, Y.; Fan, L.; Liu, Y.; Huang, Z.; Chen, Y.; Wang, N.; Zhang, Z. Fully sparse fusion for 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 7217–7231.
18. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight uncertainty in neural network. In Proceedings of the International conference on machine learning. PMLR, 2015, pp. 1613–1622.
19. Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D.M. Automatic differentiation variational inference. *Journal of machine learning research* **2017**.
20. Yan, Q.; Wang, H.; Ma, Y.; Liu, Y.; Dong, W.; Woźniak, M.; Zhang, Y. Uncertainty estimation in HDR imaging with Bayesian neural networks. *Pattern Recognition* **2024**, *156*, 110802.

21. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the international conference on machine learning. PMLR, 2016, pp. 1050–1059.
22. Zhao, Y.; Tian, W.; Cheng, H. Pyramid Bayesian method for model uncertainty evaluation of semantic segmentation in autonomous driving. *Automotive Innovation* **2022**, *5*, 70–78.
23. Bethell, D.; Gerasimou, S.; Calinescu, R. Robust uncertainty quantification using conformalised Monte Carlo prediction. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 20939–20948.
24. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* **2017**, *30*.
25. Molchanov, D.; Ashukha, A.; Vetrov, D. Variational dropout sparsifies deep neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, 2017, pp. 2498–2507.
26. Peng, L.; Wang, H.; Li, J. Uncertainty evaluation of object detection algorithms for autonomous vehicles. *Automotive Innovation* **2021**, *4*, 241–252.
27. Sharma, N.K.; Saharia, M. DeepSARFlood: Rapid and automated SAR-based flood inundation mapping using vision transformer-based deep ensembles with uncertainty estimates. *Science of Remote Sensing* **2025**, *11*, 100203.
28. Zou, K.; Yuan, X.; Shen, X.; Chen, Y.; Wang, M.; Goh, R.S.M.; Liu, Y.; Fu, H. EvidenceCap: Towards trustworthy medical image segmentation via evidential identity cap. *arXiv preprint arXiv:2301.00349* **2023**.
29. Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051* **2021**.
30. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics yolov8. 2023. URL <https://github.com/ultralytics/ultralytics> **2023**.
31. Simony, M.; Milzy, S.; Amendey, K.; Gross, H.M. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In Proceedings of the Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.
32. Shafer, G. Dempster-shafer theory. *Encyclopedia of artificial intelligence* **1992**, *1*, 330–331.
33. Jsang, A. Subjective Logic: A formalism for reasoning under uncertainty **2018**.
34. Sentz, K.; Ferson, S. Combination of evidence in Dempster-Shafer theory **2002**.
35. Jøsang, A.; Hankin, R. Interpretation and fusion of hyper opinions in subjective logic. In Proceedings of the 2012 15th International Conference on Information Fusion. IEEE, 2012, pp. 1225–1232.
36. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 3354–3361.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.