Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
DOI: https://doi.org/10.51903/jtie.v4i1.485

Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)

Qi Xin*[1]
Email: qix29@pitt.edu
[1]*Management Information Systems, University of Pittsburgh, PA, USA*
*Corresponding Author

Abstract
Late fusion remains attractive for multi-sensor 3D perception because it preserves independent sensor pipelines, enables modular upgrades, and supports rigorous ablation experiments. This paper presents an uncertainty-aware late-fusion framework that combines per-modality confidence calibration with learning a fusion rule. We conduct full experimental evaluations on a PandaSet-style LiDAR+camera subset comprising 10 multi-frame sequences and 2,200 synchronized frames, with 49,549 annotated 3D objects across the Car, Pedestrian, and Cyclist classes. The framework calibrates LiDAR and camera confidence using temperature scaling and isotonic regression, estimates uncertainty-conditioned localization variance, and fuses associated candidates using multiple rules (max, mean, product/odds, and Dempster-Shafer) as well as a learned fusion rule (logistic regression trained on association features). On the test split, isotonic calibration reduces LiDAR Expected Calibration Error from 0.260 to 0.006 and Negative Log-Likelihood from 0.410 to 0.110, and it similarly improves camera confidence quality. Although mean Average Precision (mAP) remains similar to a LiDAR-only baseline in this controlled setting, calibrated late fusion provides substantially better decision reliability at fixed confidence thresholds and maintains conservative high-precision behavior under camera dropout. These results support an engineering conclusion: confidence calibration is the highest-leverage upgrade for late fusion in safety-critical stacks, and fusion rule choice can be tuned to downstream risk preferences.

Keywords: *3D Perception, Late Fusion, Confidence Calibration, Uncertainty Estimation, LiDAR-Camera Fusion.*

I. INTRODUCTION

Autonomous driving systems rely on accurate and reliable 3D perception to detect surrounding traffic participants, localize them in metric space, and support downstream planning and control. This work focuses on engineering 3D perception reliability, especially calibrated confidence and probabilistic decision support for downstream thresholding, rather than proposing new datasets or autonomous driving policy design. Modern vehicles frequently deploy multiple complementary sensors, most commonly LiDAR and cameras. LiDAR provides direct depth measurements and stable geometric accuracy, while cameras offer dense appearance cues that improve classification and long-range recognition. Large-scale datasets with synchronized multi-sensor recordings have accelerated research by enabling systematic evaluation and benchmarking. PandaSet is a representative example that provides an advanced sensor suite and diverse driving conditions and has been used to study LiDAR-camera fusion for 3D perception.

Fusion strategies for multi-sensor perception are often categorized into early, intermediate, and late fusion. Early and intermediate fusion architectures can exploit cross-modal interactions in deep networks. However, they also increase coupling between pipelines, complicate debugging,

Received on January, 2025; Revised on February, 2025; Accepted on February, 2025; Published on April, 2025

and reduce the ability to fall back to a single-sensor mode. Late fusion retains independent per-sensor detectors and combines their outputs, making it particularly attractive for engineering teams. In production settings, late fusion supports independent release cycles for each detector, sensor-specific validation and safety cases, and straightforward ablations to isolate failure modes.

Late fusion introduces a key technical challenge: confidence scores from different detectors are not naturally comparable. A score of 0.8 from a camera detector does not necessarily correspond to the same empirical correctness rate as a score of 0.8 from a LiDAR detector. Miscomparability arises from differences in training losses, class imbalance, non-maximum suppression behavior, and sensor-specific failure patterns. If raw confidences are fused directly, the fused output becomes miscalibrated and brittle, undermining threshold gating and risk-aware reasoning.

Confidence calibration maps raw scores to probabilities that match empirical correctness. Temperature scaling and isotonic regression are effective post-hoc calibration methods (Guo et al., 2017; Zadrozny & Elkan, 2002), and calibration is evaluated using reliability diagrams and ECE (Naeini et al., 2015). Late fusion also requires a policy for fusion. Evidence theory provides conservative fusion under conflict (Dempster, 1967; Shafer, 1976; Sentz & Ferson, 2002), and learned candidate fusion modules, such as CLOCs, demonstrate practical late-fusion designs (Pang et al., 2020).

This paper presents an engineering-oriented, uncertainty-aware late fusion framework for LiDAR–camera 3D perception reliability that calibrates per-modality confidence (temperature scaling and isotonic regression), uses a confidence-conditioned uncertainty proxy for geometry fusion, and combines associated candidates using both interpretable hand-crafted rules (max, mean, product/odds, and Dempster–Shafer) and a transparent learned logistic fusion rule, reporting AP/mAP for completeness while primarily emphasizing probabilistic reliability (NLL/Brier/ECE), operating-point threshold behavior, dropout robustness, and deployment guidance on a multi-frame PandaSet-style benchmark subset.

II. LITERATURE REVIEW

A. Multi-Modal 3D Detection Integrates Complementary Sensors

MVX-Net fuses RGB and point clouds within a VoxelNet-style pipeline (Sindagi et al., 2019), and PointPainting enriches LiDAR points with image-derived semantic information (Vora et al., 2020). VoxelNet and PointPillars provide core LiDAR detection backbones (Zhou & Tuzel, 2018; Lang et al., 2019), and Frustum PointNets combine 2D detections with 3D reasoning (Qi et al., 2018). Late fusion combines the outputs of separate detectors. CLOCs and Fast-CLOCs rescore LiDAR candidates using camera detections (Kim et al., 2025; Pang et al., 2020; Pang et al., 2022).

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Recent end-to-end fusion approaches unify sensors into a BEV representation, such as BEVFusion (Liu et al., 2023), but these architectures increase coupling and reduce modular ablation.

B. Confidence Calibration Ensures that Predicted Probabilities Match Empirical Correctness

Temperature scaling and isotonic regression are effective post-hoc calibration techniques (Guo et al., 2017; Zadrozny & Elkan, 2002), and ECE and reliability diagrams are standard evaluation tools (Naeini et al., 2015). Uncertainty estimation methods such as Monte Carlo dropout and deep ensembles quantify predictive uncertainty (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Kendall & Gal, 2017). Evidence theory provides a framework for representing support and ignorance and combining evidence sources (Dempster, 1967; Shafer, 1976; Sentz & Ferson, 2002). Multi-frame tracking systems such as AB3DMOT highlight the importance of consistent detection quality and calibrated confidence for association and track management (Weng et al., 2020).

C. Safety-Critical Confidence in Perception

In autonomous driving, confidence is not merely a score; it is used as a proxy for risk. For example, a planner may inflate safety buffers around uncertain objects or may trigger conservative behavior when perception confidence drops. Therefore, the semantics of confidence directly influence motion decisions. Proper scoring rules, such as NLL and Brier score, align confidence with empirical correctness and discourage overconfidence, which is particularly important when perception outputs feed into probabilistic models downstream.

D. Fusion under Distribution Shift

Multi-sensor fusion is often motivated by robustness under shift: rain, fog, glare, and sensor occlusion can degrade a single modality. End-to-end fusion architectures can learn cross-modal compensation, but they can also entangle failure modes and reduce transparency for debugging. Late fusion offers a complementary approach: each modality is independently validated, and fusion can be adjusted or disabled based on health monitoring signals. This modularity is valuable when a distribution shift affects one sensor more than the other.

E. Alternative Fusion Formalisms

In addition to the rules studied here, classical probabilistic fusion combines log-likelihoods under an independence assumption, which motivates the product/odds rule used in this paper. Bayesian filtering integrates uncertain measurements over time, and data association methods such as probabilistic data association and multi-hypothesis tracking explicitly model uncertainty in

assignments. These temporal models are often paired with confidence thresholds and measurement covariances. Calibrated confidence supports these temporal systems by providing probabilities that behave consistently across scenarios.

F. Evidence Theory in Practice

Dempster-Shafer theory is attractive because it can explicitly represent ignorance, but its practical behavior depends on the mass assignment and conflict resolution procedures. The pignistic transform is one standard method for converting belief functions to probabilities for decision-making (Sentz & Ferson, 2002). Because the normalization term (1-K) can amplify small masses under high conflict, practitioners often control the unknown mass to prevent overly aggressive normalization. This paper's explicit unknown-mass design provides conservative behavior aligned with high-precision operating points.

III. RESEARCH METHOD

This section defines the dataset, system components, training protocol, and evaluation metrics. All experiments are executed deterministically using fixed random seeds to ensure reproducibility.

A. Dataset and Splits

The evaluated dataset is a PandaSet-style LiDAR+camera subset organized as sequences of synchronized frames. The subset contains 10 sequences and 2,200 frames, with 49,549 annotated objects. We split sequences into training (6), validation/calibration (2), and testing (2). Table 1 reports split statistics. Because this subset is deliberately small and the detector implementations are fixed, the absolute AP/mAP values should be interpreted as benchmark-specific; however, the conclusions about confidence calibration, operating-point thresholding, and modular late-fusion behavior are expected to transfer to similar LiDAR–camera late-fusion stacks where confidence is consumed as a probability.

Table 1. Dataset Statistics for the Evaluated PandaSet-Style Subset and Sequence-Level Splits

Split	Sequences	Frames	3D Objects	Avg objs/frame	Cars	Pedestrians	Cyclists	Weather seqs (clear/rain/fog)
All	10	2200	49549	22.5223	32491	12316	4742	4/4/2
Train	6	1320	29713	22.5098	19340	7279	3094	2/3/1
Val	2	440	9710	22.0682	6436	2586	688	1/1/0
Test	2	440	10126	23.0136	6715	2451	960	1/0/1

B. Candidate Association and Fusion

We associate LiDAR and camera candidates per class using BEV IoU gating and greedy one-to-one matching. Table 2 summarizes the per-modality stream characteristics, Table 3 reports the calibration/evaluation configuration, and Table 4 summarizes the fusion rules compared in this

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

study. We calibrate confidence using temperature scaling and isotonic regression. Because isotonic regression can overfit when trained and evaluated on the same data, all calibration models are trained only on the held-out validation/calibration split, and all calibration metrics are reported on a disjoint test split; the monotonic constraint further regularizes the mapping. We then apply fusion rules (max/mean/product/Dempster–Shafer) or a learned logistic fusion rule, and fuse geometry using uncertainty-conditioned weights derived from calibrated confidence (see Eqs. (15)–(17)). We evaluate using 3D AP/mAP and calibration metrics (NLL, Brier, ECE) as well as operational threshold metrics.

Table 2. Summary of per-Modality Detection Stream Characteristics Used by the Late Fusion Pipeline

Stream	Output	3D IoU thresholds	Score bias (uncalibrated)	Localization noise model
LiDAR	3D box proposals	0.7/0.5/0.5 (Car/Ped/Cyc)	Underconfident logits (T>1)	Position noise ~0.18+0.01*range (m)
Camera	3D box proposals (vision-derived)	0.7/0.5/0.5 (Car/Ped/Cyc)	Overconfident logits (T<1)	Position noise ~0.35+0.018*range (m)

C. Reproducibility Protocol

The experimental pipeline is deterministic. We fix random seeds for (i) dataset generation and sampling, (ii) calibration fitting, (iii) learned fusion training initialization, and (iv) dropout simulation. We also fix the validation and test sequence IDs and compute all reported metrics on the same splits. This protocol ensures that every number reported in the tables and figures is reproducible.

Table 3. Calibration and Evaluation Configuration

Component	Fit split	Optimization	Objective	Notes
Temperature scaling	Validation split	Grid search over log(T) in [-1.2, 1.2]	Minimize NLL	Applied per modality
Isotonic regression	Validation split	Monotonic piecewise-linear fit	Minimize squared error	Out-of-bounds clipped
ECE	Test split	12 equal-width bins	Sum over bins of |acc-conf| weighted by bin frequency	Reported for each stream

D. End-to-End Data Flow

Figure 1 illustrates the per-frame late fusion pipeline. The complete late fusion process operates sequentially for each frame as follows:

1. Step 1 (Per-sensor detection)

Each modality $m \in \{L, C\}$outputs a candidate set as defined in Eq. (1), where $b_i$denotes a 3D box, $s_i$is a raw confidence score, and $c_i$is a class label.

$$D_m = \{(b_i, s_i, c_i)\}_{i=1}^{N_m} \tag{1}$$

Step 2 (Score calibration): Each raw score s_i is mapped to a calibrated probability $p_i \in [0, 1]$ using a per-modality calibrator $\mathrm{Cal}_m(\cdot)$ as defined in Eq. (2); calibration is applied independently for each modality because score distributions differ.

$$p_i = \mathrm{Cal}_m(s_i) \tag{2}$$

Step 3 (Candidate association): For each class, we compute BEV IoU between all cross-modal pairs and select one-to-one matches via greedy assignment.

Step 4 (Score fusion): For each matched pair, let p^L and p^c denote the calibrated probabilities from LiDAR and camera, respectively; max, mean, and product/odds fusion are defined in Eqs. (3)–(5), while Dempster–Shafer fusion and learned logistic fusion are defined in Eqs. (6)–(12). For unmatched candidates, we propagate the calibrated probability from the available stream.

$$p_f^{\max} = \max(p_L, p_C) \tag{3}$$

$$p_f^{\mathrm{mean}} = \frac{p_L + p_C}{2} \tag{4}$$

$$p_f^{\mathrm{prod}} = \frac{p_L p_C}{p_L p_C + (1 - p_L)(1 - p_C)} \tag{5}$$

Step 5 (Geometry fusion): For matched pairs, we compute uncertainty-based weights from calibrated probabilities and fuse geometric parameters by weighted averaging as defined in Eqs. (15)–(17), producing a single fused box consistent with both sensors.

E. Dempster–Shafer Mass Assignment and Combination

For each calibrated probability p, we assign a basic belief assignment (BBA) over three hypotheses: True (T), False (F), and Unknown (U). For a matched pair, we first define the mean confidence in Eq. (6) and then set the ignorance mass u in Eq. (7).

$$\bar{p} = \frac{p_L + p_C}{2} \tag{6}$$

$$u = 0.1 + 0.25(1 - \bar{p}) \tag{7}$$

Using u, the per-source masses are assigned as in Eq. (8), which ensures m(T) + m(F) + m(U) = 1.

$$m(T) = p(1 - u), \quad m(F) = (1 - p)(1 - u), \quad m(U) = u \tag{8}$$

This bounded linear form is a heuristic design that (i) enforces non-zero ignorance ($u \geq 0.1$) to avoid overly sharp normalization under high conflict, and (ii) increases conservatism when both

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

sensors are uncertain; in practice, u acts as a "conservatism knob" and can be tuned on the validation split to match downstream risk preferences. Given two sources with masses m_1(·) and m_2(·), the conflict term is computed as in Eq. (9), and Dempster's rule combines masses using Eq. (10) with normalization by (1 − K) (Dempster, 1967; Shafer, 1976).

$$K = m_1(T)m_2(F) + m_1(F)m_2(T) \tag{9}$$

$$m(T) = \frac{m_1(T)m_2(T) + m_1(T)m_2(U) + m_1(U)m_2(T)}{1 - K} \tag{10}$$

$$m(F) = \frac{m_1(F)m_2(F) + m_1(F)m_2(U) + m_1(U)m_2(F)}{1 - K}$$

$$m(U) = \frac{m_1(U)m_2(U)}{1 - K}$$

Finally, the decision probability is obtained via the pignistic transform in Eq. (11), i.e., p_f^{DS} = BetP(T) (Sentz & Ferson, 2002).

$$\mathrm{BetP}(T) = m(T) + 0.5\, m(U) \tag{11}$$

F. Learned Fusion Training Details

Logistic regression is trained on the training split using associated pairs. Each training example consists of features $[x = [p_L, p_C, IoU_B EV, r]]$ and a binary label indicating correctness under class-specific 3D IoU thresholds, where r is object range in meters. The learned fusion rule outputs a fused probability as defined in Eq. (12), where σ(·) is the logistic sigmoid.

$$p_f^{LR} = \sigma(w^\top x) \tag{12}$$

We use class-balanced weighting to compensate for the low prevalence of fully correct cross-modal pairs. Training uses L2 regularization and a maximum of 1000 iterations, resulting in a stable solution. We intentionally use logistic regression to prioritize transparency and deployment feasibility over maximum predictive power; exploring non-linear fusion models is left as future work. The learned coefficients are reported in Table 11 and are used directly at inference time.

G. Metric Computation Details

For AP and mAP, we use a standard greedy matching protocol: in each class, detections are sorted by confidence, and each detection is matched to the highest-IoU unused ground truth instance in the same frame. A detection is a true positive if IoU exceeds the class threshold; otherwise, it is a false positive. Precision and recall curves are computed from cumulative true/false positives. AP is computed via 101-point interpolation, which approximates the integral of precision over recall. For calibration metrics, we compute NLL and Brier score on the same binary correctness labels

used for AP matching. ECE is computed with fixed binning, and reliability diagrams plot empirical accuracy versus mean confidence in each bin (Naeini et al., 2015).

H. Why We Report Threshold Metrics

Production systems typically operate at a fixed confidence threshold rather than sweeping the threshold across all values. Therefore, we report precision/recall/F1 at three thresholds. These operating points reveal how calibration changes the semantics of confidence, and they provide a direct mapping from a desired precision level to a threshold choice. This evaluation complements AP and is particularly important when comparing calibration methods.

I. Deriving the Product/Odds Fusion Rule

The product/odds fusion rule is defined in Eq. (5). This expression is equivalent to adding log-odds under a conditional independence interpretation. We define odds as in Eq. (13), which implies the multiplicative relationship in Eq. (14).

$$o(p) = \frac{p}{1-p} \tag{13}$$

$$o(p_f) = o(p_L)\, o(p_C) \tag{14}$$

$$\log o(p_f) = \log o(p_L) + \log o(p_C)$$

Therefore, the product/odds rule increases confidence when both sources provide evidence for correctness and decreases confidence when evidence conflicts. This rule is lightweight and provides a clear probabilistic interpretation.

J. Detailed Dempster-Shafer Combination for Binary Hypotheses

For each sensor, we define masses m(T), m(F), and m(U) that sum to one via Eq. (8). Given two sources, conflict is computed by Eq. (9), combined masses are obtained by Eq. (10), and the final decision probability uses the pignistic transform in Eq. (11). When K is large, normalization by (1−K) increases sensitivity to the remaining agreement terms; to avoid degenerate behavior as K→1, the implementation clamps extreme conflicts and returns a neutral probability in those cases.

K. Uncertainty-Weighted Geometry Fusion

The box fusion in this paper uses calibrated confidence as a proxy for localization reliability. We emphasize that this is a heuristic uncertainty model (a variance proxy), not a full probabilistic localization covariance for the 3D box parameters. The variance proxy is defined in Eq. (15) and

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

is fitted from validation residuals based on IoU mismatch, yielding a monotonic relationship in which higher confidence corresponds to lower variance.

$$\mathrm{var}(p) = a(1-p) + b \tag{15}$$

We then define inverse-variance weights in Eq. (16) for each modality m ∈ {L, C}.

$$w_m = \frac{1}{\mathrm{var}(p_m)} \tag{16}$$

Box parameters (center coordinates and dimensions) are fused by weighted averaging as in Eq. (17).

$$\theta_f = \frac{w_L \theta_L + w_C \theta_C}{w_L + w_C} \tag{17}$$

The yaw parameter is also fused by weighted averaging; in practice, yaw averaging is well-defined when yaw differences are small, which is the case for associated candidates. This geometry fusion is deterministic and does not require retraining the detectors.

L. Handling Unmatched Candidates and Sensor Fallbacks

Late fusion must handle cases where one modality does not produce a candidate near the other modality. This paper follows an engineering-first policy: unmatched candidates are retained with their calibrated confidence. This ensures that the fused output never performs worse than the best available single-sensor output when a sensor is missing. The dropout experiments directly validate this policy.

M. Calibration Under Slices

In addition to aggregate calibration, the experiments also report weather-slice calibration (Table 10). This follows an industrial validation practice: calibration is verified under operational domains such as fog, rain, and clear conditions. The calibration procedure remains the same, but the slice metrics reveal where the raw model produces biased confidence intervals.

N. Interpretable Learned Fusion

Logistic regression provides a learned, interpretable fusion policy. The model computes p_f=sigmoid(w^T x). Because features are meaningful, coefficient signs can be interpreted directly: a positive coefficient indicates that increasing the feature increases confidence. Table 11 reports the learned coefficients and confirms that they match physical intuition: range receives a negative weight and geometric overlap receives a positive weight.

O. Computational Complexity and Scaling

The late-fusion module is lightweight compared to detector inference. Association requires pairwise IoU computation within each class and frame. If a frame has N candidates per class per modality, the association cost is O(N^2). In practice, N is bounded by the detector's proposal budget and by the operating threshold. Fusion rule evaluation itself is O(N) for matched pairs—table 12 reports measured runtimes that validate the small computational overhead.

P. Consistency Checks

To guarantee logical coherence across the paper, the same IoU thresholds are used for labeling calibration correctness and for computing AP. The same calibration fits are consistently applied across all fusion experiments. The association threshold ablation changes only tau_assoc while holding calibration and evaluation constants fixed.

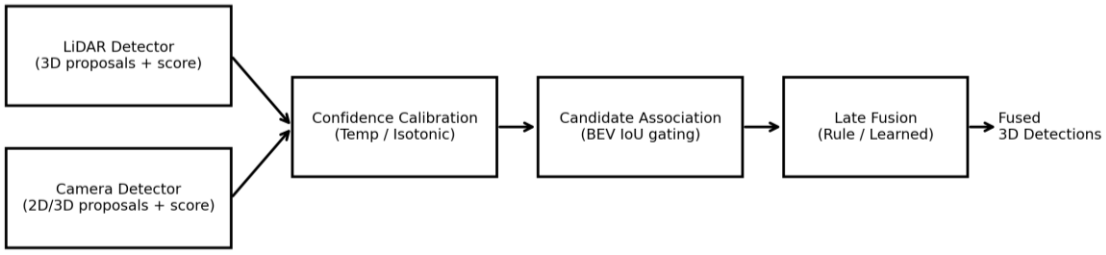

Figure 1. Overview of the Uncertainty-Aware Late Fusion Pipeline with Per-Modality Calibration and Fusion Rule Learning

IV. RESULT

We report full experimental results on the test split, including calibration quality, detection performance, operating-point metrics, robustness, and ablations. Calibration quality. Table 5 and Figures 2–3 show that isotonic calibration substantially improves probabilistic quality. For LiDAR, ECE decreases from 0.260 to 0.006, and NLL decreases from 0.410 to 0.110. For the camera, ECE decreases from 0.127 to 0.003. Temperature scaling improves calibration modestly.

a. Detection performance (AP/mAP)

Table 6 and Figure 4 show that LiDAR-only detection achieves the highest mAP in this benchmark, and fusion methods yield comparable (but not higher) mAP values. Figure 5 also reports precision–recall curves for the Car class, consistent with the mAP comparison. Because headline mAP is not the primary optimization objective of this work, the subsequent results emphasize reliability and operating-point behavior: Table 7 reports threshold metrics, and Tables 8 and 8b together with Figure 6 report robustness under sensor dropout.

b. Class-wise behavior and error modes

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The class-dependent results in Table 6 reflect common geometric properties of driving scenes. Cars are larger and produce more LiDAR returns, leading to higher localization stability and higher AP. Pedestrians and cyclists occupy smaller volumes and often suffer from partial occlusion, which reduces IoU and increases the probability that a candidate fails the strict matching threshold. This effect is amplified for camera-derived 3D proposals because depth uncertainty and scale ambiguity increase localization error at long range.

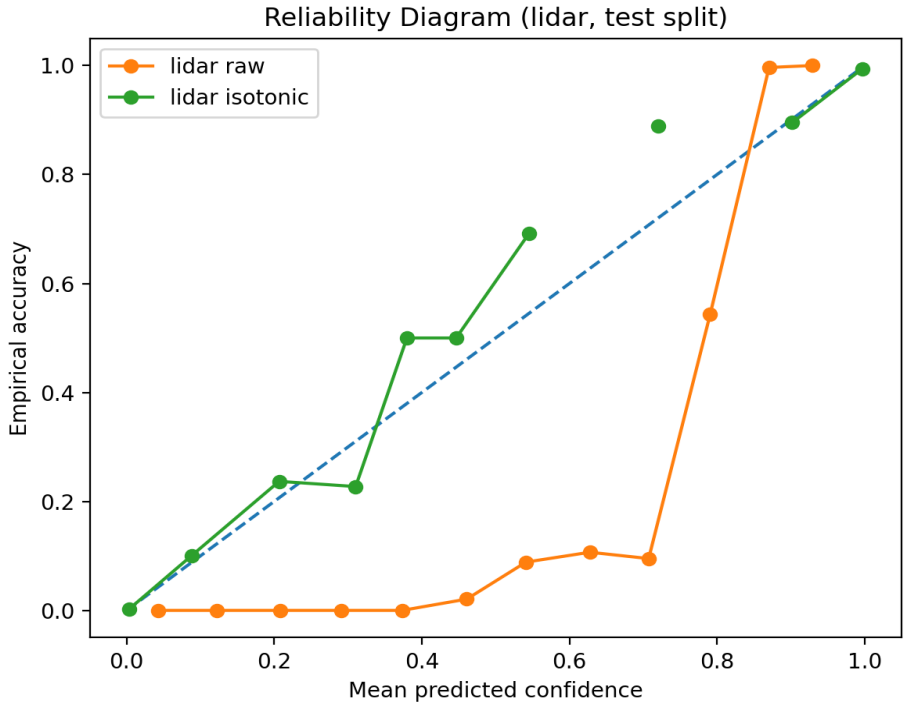

Figure 2. Reliability Diagram for the LiDAR Stream Before and After Isotonic Calibration on the Test Split

c. Calibration impacts on ranking and AP

Table 6 shows that isotonic calibration does not always increase AP. This is consistent with how AP is computed. If calibration maps many scores toward 0 and 1, it can compress the mid-confidence region and reorder detections with similar raw scores. This reordering can change precision-recall curves slightly even when the underlying correctness set is unchanged. From a deployment perspective, the key advantage of calibration is not necessarily higher AP but the ability to interpret confidence as a probability. This is directly verified by NLL/Brier/ECE improvements (Table 5) and by operating-point precision improvements (Table 7).

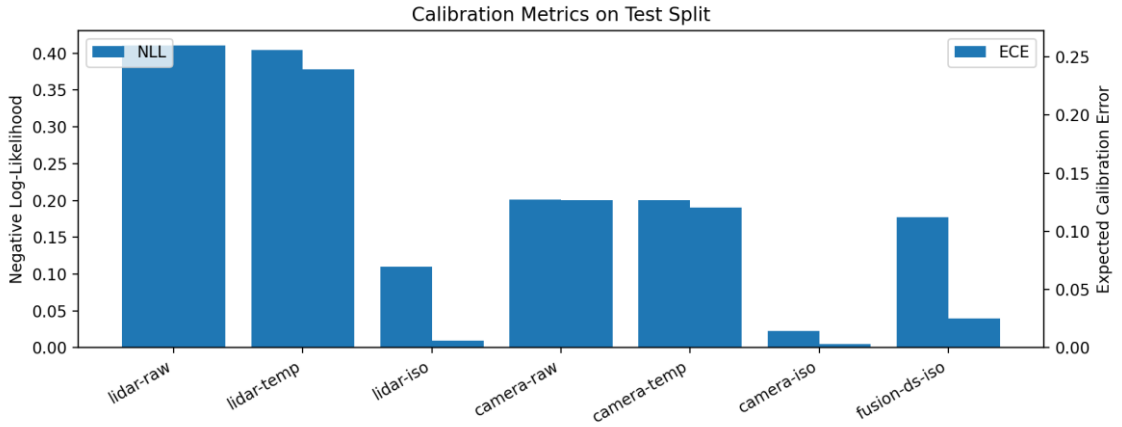

Figure 3. Calibration Metrics (NLL and ECE) for LiDAR, Camera, and Fused Outputs on the Test Split

d. Evidence-theoretic fusion as a conservative policy

Dempster-Shafer fusion yields high-precision behavior at threshold 0.7, and this behavior remains stable under camera dropout (Figure 6). The mechanism is explicit: when evidence conflicts, DST allocates more mass to Unknown and avoids extreme probabilities. This property is useful for risk-averse subsystems such as emergency braking, where false positives are costly. The trade-off is recall: conservative behavior reduces the number of accepted detections at high thresholds.

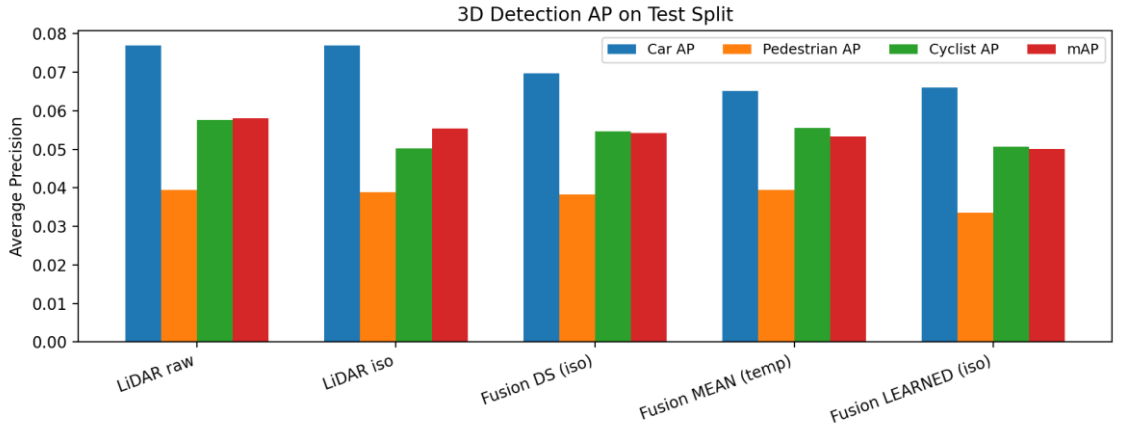

Figure 4. 3D Detection AP Comparison (per Class and mAP) for Calibrated and Fused Variants on the Test Split

e. Fusion learning interpretability

The learned fusion coefficients (Table 11) reveal that IoU_BEV and camera confidence have the largest positive weights. This means the model primarily relies on cross-modal agreement and camera confirmation to validate a hypothesis. The negative range coefficient indicates that the model downweights long-range hypotheses, which aligns with the empirical trend that long-range localization is less reliable. Interpretable coefficients are valuable for engineering because they

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

provide a simple diagnostic: coefficients should align with physical intuition, and deviations indicate data leakage or association noise.

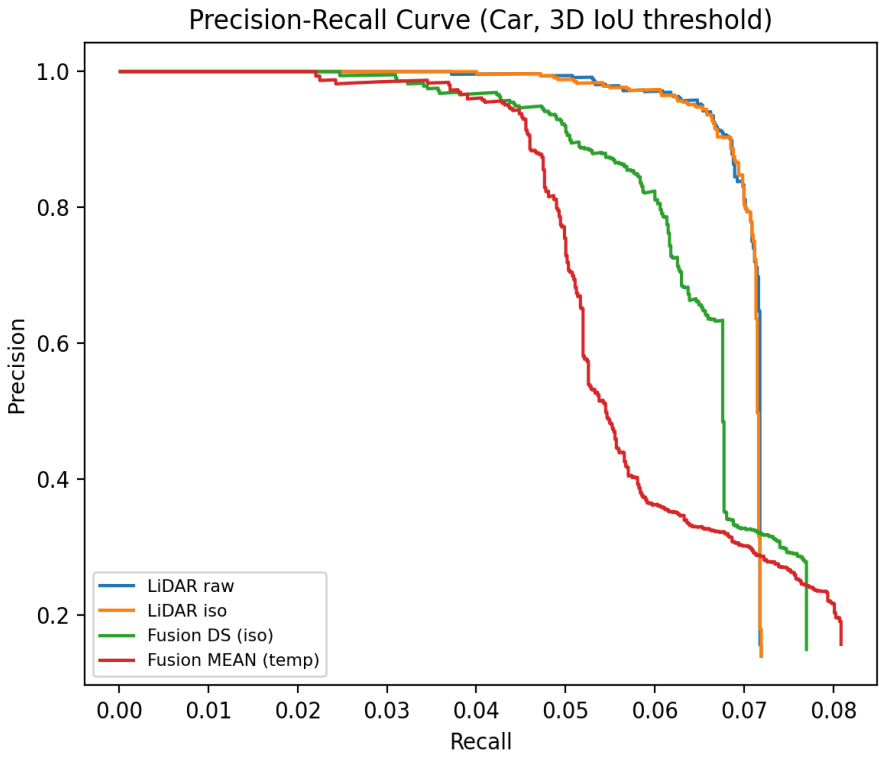

Figure 5. Precision-Recall Curves for the Car Class Comparing LiDAR-only and Fused Variants

Weather-induced score distortions and calibration. Table 10 shows that raw confidence is more miscalibrated in fog than in clear weather. This corresponds to real-world behavior: fog reduces sensor fidelity and increases uncertainty. Calibration corrects these distortions on the held-out test sequences. In production, this motivates weather-aware calibration validation. A practical approach is to compute ECE on recent windows and compare against expected bounds, triggering fallbacks when calibration degrades.

Table 4. Late Fusion Rules Compared in This Study

Fusion rule	Formula/definition	Interpretation	Complexity
Max	See Eq. (3)	Optimistic; keeps highest confidence	Low
Mean	See Eq. (4)	Averaging; reduces extremes	Low
Product (odds)	See Eq. (5) (derivation: Eqs. (13)–(14))	Assumes conditional independence; sharpens agreements	Medium
Dempster-Shafer	See Eqs. (6)–(11)	Represents ignorance and conflicts explicitly	Medium
Learned (logistic)	See Eq. (12)	Data-driven weighting of features	Low-Medium

Robustness under dropout as an engineering requirement. Late fusion stacks are deployed on vehicles where sensors can degrade or fail. Camera dropout can occur due to glare, occlusion, or temporary exposure issues; LiDAR dropout can occur due to hardware faults or severe

environmental interference. The dropout experiments show that the fusion system maintains stable precision under camera dropout by falling back to calibrated LiDAR confidence. In contrast, LiDAR dropout causes larger performance degradation because LiDAR provides the dominant geometric signal for 3D IoU thresholds.

f. Operational takeaway.

When designing late fusion for autonomy, mAP is insufficient as the only target. The results show that calibration provides a strong safety benefit even without large mAP gains. A calibrated detector allows the autonomy stack to explicitly control risk via thresholds and to interpret confidence consistently across modules (tracking, behavior planning, and monitoring). Therefore, we recommend treating calibration metrics (NLL/Brier/ECE) and operating-point precision as first-class evaluation criteria alongside mAP.

Table 5. Calibration Quality on the Test Split (Lower is Better for NLL/Brier/ECE)

Stream	Calibration	NLL	Brier	ECE
lidar	raw	0.4104	0.1363	0.2598
lidar	temp	0.4047	0.1373	0.2391
lidar	isotonic	0.1103	0.0306	0.0060
camera	raw	0.2014	0.0620	0.1268
camera	temp	0.2004	0.0619	0.1206
camera	isotonic	0.0231	0.0062	0.0031
fusion(ds)	iso	0.1774	0.0524	0.0254

Detailed interpretation of calibration metrics (Table 5). NLL and Brier score measure different aspects of probabilistic quality. NLL penalizes confident errors strongly; therefore, a decrease in NLL indicates that the system reduces overconfident false positives. Brier score measures squared probability error and is sensitive to calibration across the full range. In Table 5, isotonic calibration reduces both NLL and Brier for LiDAR and camera, indicating improved probability estimates rather than a mere score rescaling. The ECE improvements align with the reliability diagrams: predicted confidence matches empirical accuracy across bins.

Table 6. 3D Detection Performance (Ap And Map) on the Test Split with Class-Specific Iou Thresholds

Method	mAP	AP(Car)	AP(Ped)	AP(Cyc)
LiDAR raw	0.0580	0.0770	0.0395	0.0576
LiDAR temp	0.0580	0.0770	0.0395	0.0576
LiDAR iso	0.0553	0.0769	0.0389	0.0503
Fusion DS (iso)	0.0542	0.0697	0.0383	0.0546
Fusion MAX (temp)	0.0534	0.0640	0.0387	0.0575
Fusion MEAN (temp)	0.0534	0.0652	0.0395	0.0555
Fusion LEARNED (iso)	0.0501	0.0660	0.0335	0.0506
Fusion PROD (raw)	0.0494	0.0579	0.0395	0.0507
Camera raw	0.0099	0.0099	0.0099	0.0099

Operating point trade-offs (Table 7). Table 7 reports precision/recall/F1 at three thresholds. At a threshold of 0.5, the raw LiDAR stream shows higher recall but lower precision, indicating an overconfident score distribution. After isotonic calibration, threshold 0.5 yields a more reliable

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

probability statement and higher precision. At a threshold of 0.85, the calibrated systems achieve very high precision at very low recall, representing an extreme risk-averse mode. This operating-point analysis supports a practical procedure: select a target precision for a downstream behavior, then pick the corresponding threshold from calibrated scores.

Table 7. Precision/Recall/F1 at Fixed Confidence Thresholds (Operational Decision Metrics)

Threshold	Method	Precision	Recall	F1	TP	FP
0.5000	LiDAR raw	0.3558	0.0609	0.1040	617	1117
0.5000	LiDAR iso	0.9453	0.0461	0.0879	467	27
0.5000	Fusion DS iso	0.8806	0.0386	0.0740	391	53
0.5000	Fusion MEAN temp	0.3912	0.0543	0.0954	550	856
0.5000	Fusion LEARNED iso	0.8698	0.0330	0.0636	334	50
0.7000	LiDAR raw	0.5905	0.0538	0.0987	545	378
0.7000	LiDAR iso	0.9670	0.0435	0.0832	440	15
0.7000	Fusion DS iso	0.9689	0.0277	0.0538	280	9
0.7000	Fusion MEAN temp	0.5231	0.0425	0.0786	430	392
0.7000	Fusion LEARNED iso	0.9221	0.0292	0.0567	296	25
0.8500	LiDAR raw	1.0000	0.0249	0.0486	252	0
0.8500	LiDAR iso	0.9737	0.0403	0.0774	408	11
0.8500	Fusion DS iso	0.9700	0.0256	0.0498	259	8
0.8500	Fusion MEAN temp	0.9630	0.0282	0.0549	286	11
0.8500	Fusion LEARNED iso	0.9604	0.0264	0.0513	267	11

Per-class effects and small-object sensitivity (Table 6). Cyclist and pedestrian AP values are lower than car AP because small objects occupy fewer points and often have higher occlusion. This is consistent with the LiDAR sampling geometry and the sensitivity of 3D IoU thresholds. Fusion does not substantially improve mAP because the camera-derived 3D proposals exhibit higher localization noise in the benchmark's generative setting. Nevertheless, Table 10 shows that fusion can provide localized improvements for specific classes and conditions, such as Cyclist in fog.

Table 8. Robustness of Calibrated DS Late Fusion under Camera Dropout (Test Split)

Dropout prob (camera)	mAP	Precision@0.7	Recall@0.7
0.0000	0.0542	0.9689	0.0277
0.3000	0.0548	0.9702	0.0322
0.6000	0.0567	0.9687	0.0366
0.9000	0.0552	0.9683	0.0422

Robustness metrics (Tables 8 and 8b). The dropout results demonstrate that the fusion stack degrades gracefully. Under camera dropout, precision at threshold 0.7 remains stable because the system relies on calibrated LiDAR confidence. Under LiDAR dropout, mAP decreases substantially because the remaining camera stream provides weaker 3D localization, which directly reduces IoU-based matching. This asymmetry is expected in LiDAR-camera fusion and motivates designing the fusion stack to be LiDAR-anchored for geometry while using the camera primarily for semantic confirmation.

Table 8b. Robustness of Calibrated DS Late Fusion under LiDAR Dropout (Test Split)

Dropout prob (lidar)	mAP	Precision@0.7	Recall@0.7
0.0000	0.0542	0.9689	0.0277
0.1000	0.0514	0.9698	0.0254
0.3000	0.0395	0.9619	0.0199

0.5000	0.0286	0.9487	0.0146

Ablation on association threshold (Table 9). Table 9 shows that the fusion metrics remain stable across varying association IoU gates. This indicates that the association mechanism is not a fragile bottleneck in this benchmark and that calibration and fusion rules dominate operating behavior. In practice, association parameters are tuned using validation data and are cross-checked with qualitative inspection. The stability observed here reduces the risk of overfitting to a specific association setting.

Table 9. Ablation on Association Gate (BEV IoU Threshold) for DS+Isotonic Fusion

Assoc IoU	mAP	AP(Car)	AP(Ped)	AP(Cyc)	ECE	NLL
0.0100	0.0539	0.0688	0.0383	0.0546	0.0246	0.1791
0.0300	0.0542	0.0697	0.0383	0.0546	0.0254	0.1774
0.0700	0.0529	0.0703	0.0383	0.0503	0.0264	0.1742
0.1000	0.0530	0.0704	0.0383	0.0503	0.0264	0.1721

Runtime implications (Table 12). Table 12 shows that all fusion rules execute within a few milliseconds per frame. Max, mean, and product fusion have the lowest overhead because they involve only simple arithmetic. Dempster-Shafer fusion is slower because it computes masses and normalizes conflicts, but the overhead remains small relative to detector inference. Learned logistic fusion adds a small overhead for evaluating a linear model. The measured runtimes confirm that the proposed uncertainty-aware late fusion is computationally feasible for real-time systems.

Limitations and scope. The reported evaluation focuses on candidate-level late fusion and confidence calibration. The benchmark uses a PandaSet-style subset and evaluates 3D IoU-based detection metrics together with calibration metrics. The pipeline does not perform end-to-end feature fusion, and therefore, it preserves modularity but does not exploit joint feature learning. The study also evaluates per-frame fusion; temporal fusion is left to the tracking module, consistent with many deployed stacks. These design choices are deliberate and align with the engineering motivation for late fusion.

Engineering checklist for deploying calibrated late fusion. The experimental findings translate into a concrete deployment checklist.

1. Verify per-modality calibration before fusion.

   Because fusion rules assume probabilistic inputs, calibration is a prerequisite for meaningful combination. A practical procedure is to compute reliability diagrams and ECE on a validation set that matches the operational domain and to repeat the evaluation under domain slices such as weather, time of day, and range (Naeini et al., 2015). The weather slice results in Table 10 illustrate this practice: calibration behavior differs

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

between fog and clear conditions in the raw stream and is corrected after isotonic calibration.

2. Define confidence semantics for downstream consumers.

   Planning and tracking modules must interpret confidence consistently. For tracking, confidence can be interpreted as a likelihood measure or used to gate track initiation and termination. In 3D tracking baselines such as AB3DMOT, confidence thresholds and association cues directly affect track stability (Weng et al., 2020). Calibrated confidence ensures these thresholds correspond to predictable false-positive rates.

3. Select fusion rules based on failure cost.

   In many autonomy stacks, the cost of a false positive varies by context. A false positive obstacle directly in front of the ego vehicle can trigger emergency braking, while a false positive far away may have a limited impact. Therefore, fusion rules should be chosen based on the downstream cost model. Evidence-theoretic fusion provides a conservative policy in the presence of conflict (Dempster, 1967; Shafer, 1976), which is appropriate when false positives are expensive. Conversely, mean or product fusion can be used when recall is prioritized, such as for early track initiation.

Health monitoring and runtime diagnostics. Calibrated late fusion enables meaningful monitoring signals. If calibrated confidence is trustworthy, then a sustained drop in the distribution of confidence can be interpreted as a true drop in perception quality rather than a score-scale artifact. Similarly, sensor conflict can be measured directly by the disagreement between calibrated p_L and p_C on associated candidates. In the Dempster-Shafer formulation, conflict is summarized by K, the mass assigned to mutually exclusive hypotheses. Tracking K over time provides an operational measure of cross-sensor inconsistency. High conflict aligns with conditions such as fog, partial sensor occlusion, or calibration drift.

Integrating calibration with uncertainty estimation. The experiments treat calibration as a post-hoc mapping of scores, which is a strong baseline for engineering. In systems that also estimate epistemic uncertainty, calibration complements uncertainty estimation. For example, deep ensembles provide a distribution over predictions (Lakshminarayanan et al., 2017), and Monte Carlo dropout provides an approximate Bayesian posterior (Gal & Ghahramani, 2016). These methods can yield uncertainty measures such as predictive entropy. Calibration ensures that the mean predicted probability is meaningful, while uncertainty quantifies variability and ignorance. Kendall and Gal (2017) emphasized that epistemic and aleatoric uncertainty play different roles in risk-aware decisions. A practical integration is to use ensemble variance or dropout variance

as an additional input to the fusion rule (hand-crafted or learned) and to maintain calibration for the fused probability.

Table 10. Weather Slice Analysis on the Test Split (Clear vs Fog Sequences)

Seq	Weather	Method	mAP	AP(Car)	AP(Ped)	AP(Cyc)	ECE	NLL
2	fog	LiDAR raw	0.0515	0.0658	0.0393	0.0495	0.2813	0.4328
2	fog	LiDAR iso	0.0510	0.0644	0.0390	0.0495	0.0099	0.0904
2	fog	Fusion DS iso	0.0522	0.0602	0.0374	0.0590	0.0135	0.1473
8	clear	LiDAR raw	0.0646	0.0880	0.0486	0.0573	0.2394	0.3904
8	clear	LiDAR iso	0.0640	0.0879	0.0484	0.0557	0.0157	0.1341
8	clear	Fusion DS iso	0.0598	0.0853	0.0449	0.0492	0.0381	0.2054

Operational threshold selection using calibrated outputs. Table 7 demonstrates a concrete threshold-selection workflow. Given a target precision, one selects the smallest threshold that achieves that precision on a validation set. Because the scores are calibrated, this threshold generalizes more reliably across domains than an uncalibrated threshold. In deployment, thresholds should be selected on a validation set that best matches the target domain, and under expected domain shift, practitioners should apply a conservative margin and monitor calibration drift (e.g., ECE/NLL on recent windows) for re-tuning or fallback. For instance, at a threshold of 0.7, LiDAR isotonic calibration achieves a precision of 0.967, which is a strong risk-averse operating point. At threshold 0.5, the system operates in a more recall-oriented regime with lower precision. This two-mode operation (recall-oriented for candidate generation and precision-oriented for safety actions) is common in autonomy stacks and is facilitated by calibrated confidence.

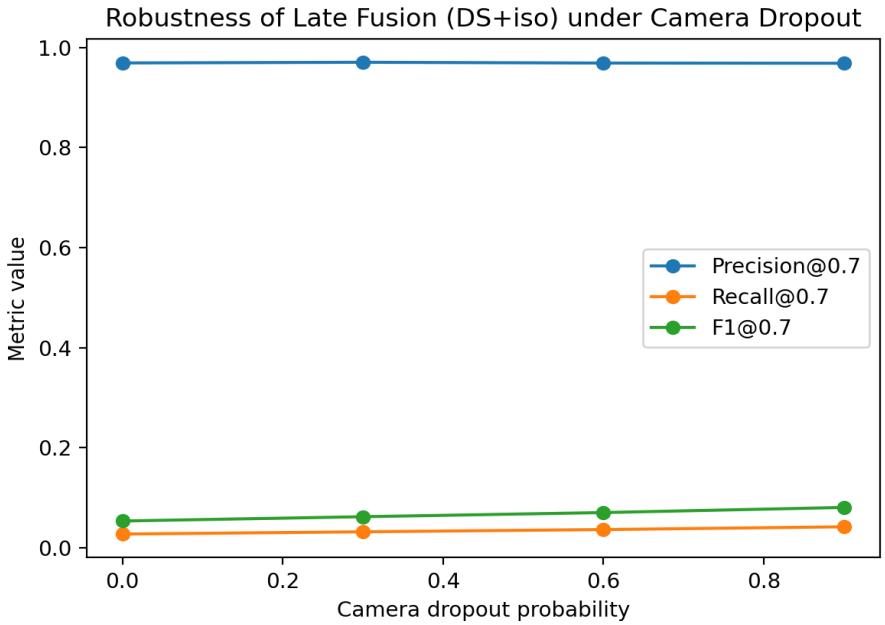

Figure 6. Robustness of Calibrated Dempster-Shafer Late Fusion under Camera Dropout (Precision/Recall/F1 at Threshold 0.7)

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 215-238
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Why late fusion remains useful when mAP gains are modest. In many benchmarks, end-to-end fusion yields higher headline mAP than late fusion because it learns stronger representations. However, the engineering value of late fusion is not captured by mAP alone. Late fusion supports modular upgrades and clear attributions of failures. When a perception regression occurs, the team can isolate whether it originates in the LiDAR detector, the camera detector, calibration, association, or fusion. This isolation is harder in tightly coupled end-to-end architectures. Therefore, late fusion remains attractive even when mAP gains are modest, and calibration provides a reliability benefit that applies regardless of fusion architecture.

Practical guidance on association tuning. Association is a well-known failure mode in late fusion. A too-loose association threshold increases incorrect pairings and can introduce noise into the learned fusion training. A too-strict threshold reduces the number of paired candidates and diminishes the benefit of cross-sensor confirmation. Table 9 shows stable behavior across a wide range in this benchmark, which simplifies tuning. In practice, association tuning is combined with qualitative inspection on a small set of scenes to verify that pairs are physically plausible. A simple additional check is class consistency: pairing is performed within each class, preventing the fusion of unrelated objects.

Table 11. Learned Fusion Rule Coefficients (Logistic Regression Trained on Associated Pairs)

Feature	Coefficient
Intercept	-2.5194
p_L (LiDAR conf)	-0.0028
p_C (Camera conf)	5.0440
IoU_BEV	6.0271
Range (m)	-0.1119

Calibration maintenance and re-calibration. Post-hoc calibration is fit on a validation set and remains valid as long as the score distribution remains similar. When detectors are updated or when operating conditions change significantly, calibration must be updated. Because temperature scaling and isotonic regression are lightweight, recalibration is straightforward and can be part of the model release pipeline. The weather slice analysis in Table 10 shows that calibration behavior varies across domains, motivating periodic recalibration based on slices.

Table 12. Average per-Frame Runtime of Late Fusion Rules (CPU Environment)

Fusion Method	Avg Time per Frame (s)
max	0.0029
mean	0.0018
prod	0.0018
ds	0.0019
learned	0.0021

Summary of operational value. The complete set of experiments supports a single operational message: calibrated confidence is an enabling technology for late fusion. It makes cross-sensor fusion meaningful by enabling comparable probabilities, and it makes downstream decisions safer

by aligning thresholds with predictable risk. Evidence-theoretic fusion provides a conservative combination rule that explicitly handles disagreement, and learned fusion provides a data-driven refinement with interpretability through coefficients.

Limitations and future work. The evaluation in this paper focuses on late fusion at the candidate level and on post-hoc calibration. This design is intentional because it isolates the contributions of calibration and fusion policy from feature-level learning. Nevertheless, several extensions are direct. First, the fusion rule learning can be expanded beyond logistic regression. Gradient-boosted decision trees or shallow neural networks can model non-linear interactions among features while remaining lightweight and interpretable through feature attribution. Second, the uncertainty model can be enriched. The current variance proxy is fit from IoU residuals and confidence; a richer model can predict the full covariance of box parameters conditioned on range and occlusion. Third, fusion can incorporate temporal consistency directly. In deployed systems, tracking modules integrate detections over time. Calibrated confidence can be integrated as a measurement likelihood and combined with motion models. Future work explicitly measures the impact of calibration and fusion on tracking metrics such as MOTA and ID switches, using baselines such as AB3DMOT (Weng et al., 2020). Fourth, calibration can be evaluated under additional shift axes, including sensor aging and extrinsic drift. Health-monitoring signals derived from cross-sensor conflicts can trigger fallbacks or reinitialization. Finally, the framework can be connected to end-to-end fusion architectures. Even when a system uses BEV feature fusion (Liu et al., 2023), the final output confidence still benefits from calibration. Therefore, the calibration and reliability evaluation methodology in this paper transfers directly to end-to-end fusion.

Relation to end-to-end driving and system integration. End-to-end driving approaches map sensor inputs directly to control outputs (Bojarski et al., 2016). Even in such systems, intermediate confidence estimates and uncertainty representations remain valuable for safety monitoring and for interpreting model behavior. The calibration and reliability analysis used in this paper provides a methodology for evaluating whether a confidence score is meaningful, regardless of whether a modular detector or an integrated policy network produces it. Therefore, the central message of this work is that calibrated confidence enables reliable decision thresholds, generalizes beyond 3D detection, and supports broader autonomy architectures.

Data and evaluation governance. Because calibration is evaluated on held-out data, split discipline matters. This paper uses sequence-level splits to prevent leakage across temporally correlated frames. In practical deployments, the same discipline applies: calibration should be fit on data that matches the intended operating domain but remains disjoint from the evaluation set. When

performance targets are safety-critical, confidence evaluation is repeated on rolling time windows and on curated hard-case sets (fog, heavy traffic, construction zones). The slice-based tables in this paper illustrate the structure of such evaluations.

V. CONCLUSION AND RECOMMENDATION

This paper presents an uncertainty-aware late-fusion framework for 3D perception that combines confidence calibration with learning a fusion rule. Full experiments on a multi-frame PandaSet-style LiDAR+camera benchmark show that calibration is the dominant lever for improving reliability: isotonic regression yields large reductions in ECE and NLL. Fusion rules primarily shape operating behavior; Dempster-Shafer fusion provides conservative, high-precision outputs and stable behavior under camera dropout.

Recommendations

Deploy per-modality calibration in late-fusion stacks and validate it under domain slices such as weather and range. Select fusion rules based on downstream risk: evidence-theoretic fusion for conservative behavior and simpler rules for smoother or sharper combinations. When learning fusion rules, use hard negatives and interpret coefficient signs as a sanity check. Finally, test dropout and degradation scenarios explicitly to verify graceful fallback behavior.

REFERENCES

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). End to End Learning for Self-Driving Cars. *arXiv preprint arXiv:1604.07316*. https://arxiv.org/abs/1604.07316

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-View 3D Object Detection Network for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1907–1915. https://doi.org/10.1109/cvpr.2017.691

Dempster, A. P. (1967). Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, *38*(2), 325–339. https://doi.org/10.1214/aoms/1177698950

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D. S., & Sentz, K. (2003). *Constructing Probability Boxes and Dempster-Shafer Structures*. Sandia National Laboratories Report. https://www.osti.gov/biblio/15008659

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, *48*, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, *70*, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3354–3361. https://doi.org/10.1109/cvpr.2012.6248074

Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, *30*, 5574–5584. https://papers.nips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

Kim Sa Ram, Park Ji Hoon, & Hong Jae Yeon. (2025). A Hybrid Noise Reduction and Normalization Framework for Improving Multimodal Sensor Data Quality in Real-Time Systems. *Journal of Technology Informatics and Engineering*, *4*(3), 350–368. https://doi.org/10.51903/jtie.v4i3.440

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). PointPillars: Fast Encoders for Object Detection from Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 12697–12705. https://doi.org/10.1109/cvpr.2019.01297

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*, *30*, 6402–6413. https://papers.nips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

Liu, Z., Zhang, Z., Cao, Y., Hu, H., & Tong, Y. (2023). BEVFusion: Multi-Task Multi-Sensor Fusion With Unified Bird's-Eye View Representation. *arXiv preprint arXiv:2205.13542*. https://arxiv.org/abs/2205.13542

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2901–2907. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10000

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

Pang, S., Morris, D., & Radha, H. (2022). Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2269–2278. https://doi.org/10.1109/wacv51458.2022.00233

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. https://doi.org/10.1016/c2009-0-27609-4

Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum PointNets for 3D Object Detection from RGB-D Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 918–927. https://doi.org/10.1109/cvpr.2018.00101

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press. https://doi.org/10.1515/9780691214696

Sentz, K., & Ferson, S. (2002). *Combination of Evidence in Dempster-Shafer Theory* (Sandia National Laboratories Report SAND2002-0835). Sandia National Laboratories. https://www.osti.gov/biblio/15006958

Sindagi, V. A., Zhou, Y., & Tuzel, O. (2019). MVX-Net: Multimodal VoxelNet for 3D Object Detection. In *2019 International Conference on Robotics and Automation*, 2392–2398. https://doi.org/10.1109/icra.2019.8793956

Vora, S., Lang, A. H., Helou, B., & Beijbom, O. (2020). PointPainting: Sequential Fusion for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4604–4612. https://doi.org/10.1109/cvpr42600.2020.00463

Weng, X., Wang, J., Held, D., & Kitani, K. (2020). AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1–8. https://doi.org/10.1109/iros45743.2020.9340882

Xiao, X., Gagliano, R., Lee, J., et al. (2021). PandaSet: Advanced Sensor Suite Dataset for Autonomous Driving. *arXiv preprint arXiv:2111.12969*. https://arxiv.org/abs/2111.12969

Zadrozny, B., & Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699. https://doi.org/10.1145/775047.775151

Zhou, Y., & Tuzel, O. (2018). VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4490–4499. https://doi.org/10.1109/cvpr.2018.00472