Data Engineering and Applications

Research Paper

Hierarchical Multimodal Data Fusion for Robust Real-Time Perception in Dynamic Sensor-Driven Environments

Szymon Wójcik[1,*] and Kacper Kaczmarek[1]

[1] Faculty of Information Engineering, Warsaw University of Technology, Warsaw, 00-661, Poland
*Corresponding author: szymon.w@pw.edu.pl

Abstract. High-accuracy, dependable perception capabilities are currently needed for more sophisticated real-time applications of autonomous robotics and intelligent surveillance systems, which are made possible by multi-modal sensor fusion. The current study investigates how to keep perception stable while dynamics, noise, and failures are present. This study proposes a multi-level fusion framework that combines several sensor data types with temporal gating, deep residual correction, and adaptive attention weighting for optimal fusion. The experiment also includes ablation research and deployments in a variety of settings, including extended real-world navigation circuits and the structured lab. The following are the comparable quantitative findings: In the event of significant sensor dropout or environmental changes, the inference delay per cycle is less than 43 ms, the mean segmentation error is less than 2.5 cm, and the top-1 identification accuracy is 94.5%. All of the architecture's modules are necessary for strong resilience and spatial precision, according to ablation experiments. The approach consistently outperforms previous baseline techniques in terms of recognition accuracy and exhibits consistent error behaviour in situations that are unfamiliar and undergoing rapid change. To put it briefly, the hierarchical fusion approach has expanded the real-time multi-modal perception state-of-the-art and offered a scalable and fault-tolerant basis for useful applications in safety-critical situations.

Keywords: *Multimodal Fusion, Real-Time Perception, Deep Learning, Sensor Robustness, Autonomous Systems*

Received on 19 August 2024, Accepted on 25 December 2024, Published on 13 January 2025

Introduction

Mobile robots are used in a variety of settings, including automated warehouses and self-driving autos [1]. The accuracy and dependability of the robots' sensing systems determine how safely and adaptably they can operate in numerous disorganised areas [2]. In order to achieve this, a variety of sensors, such as LiDAR, stereo and RGB-D cameras, radar, and inertial measurement units (IMUs), have been developed quickly in recent years [3]. Nevertheless, each modality has drawbacks of its own: IMUs can accrue drift, camera-based systems are sensitive to changes in light, and LiDAR is susceptible to being impacted by heavy rain [4]. As operational settings have grown more dynamic and complicated, the shortcomings of single-sensor systems have become more apparent, prompting the development of multi-sensor solutions [5]. In order to improve safety and perceptual reliability in real-world robot applications, research has recently been done to increase the richness and complementarity of data sources [6]. In order to create a comprehensive picture of the environment using a variety of sensors, attention has gradually shifted to the development of sophisticated data fusion techniques [7].

The primary technical issues in mobile robot perception are now thought to need multimodal data fusion frameworks [8]. By combining the geometric accuracy of LiDAR, texture and colour information from vision, and motion awareness from IMUs, the goal is to leverage the complementary capabilities of the several sensors to create a reliable and flexible scene comprehension system [9]. The extraction, representation, and fusion of

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

multi-scale spatial and semantic features from heterogeneous data have demonstrated outstanding results using state-of-the-art deep learning techniques [10]. However, there are still a number of issues that practical applications must deal with, including inadequate computational power and resources, high real-time processing demands, asynchronous data arrival from multiple sensors, and noise or misalignment in the data from these sensors [11]. The remaining issue of striking a balance between recognition accuracy and system efficiency has not been resolved in light of the aforementioned limitations [12]. Many of the current fusion solutions must be developed in a way that allows for extension under real-world operation because they are not generalisable across different contexts [13]. Therefore, in order for the next generation of mobile robots to accomplish high-performance autonomous perception in diverse and unexpected situations, a good-enough, highly efficient, and adaptive fusion mechanism is needed [14]. Research on novel multi-modal methods and systems has been driven in recent years by these technical requirements [15].

This research presents a novel and high-performance multimodal data fusion architecture for mobile robot environment perception, motivated by the aforementioned unresolved issues. The benefits of various sensors are dynamically leveraged at a high degree of efficiency and ease of implementation through the use of attention-guided feature extraction and hierarchical fusion. Experiments show that in many cases, the suggested method has achieved a high degree of recognition accuracy and decreased computational cost. The aforementioned findings offer a crucial foundation for the creation of useful, real-time multimodal perception systems for sophisticated mobile robots.

Related Work

Sensor Fusion Techniques

Sensor fusion has long been utilised to improve the robustness of the robot environment-sensing module. The Kalman filter (KF) and its extensions, such as the extended Kalman filter (EKF) and unscented Kalman filter (UKF), have been extensively utilised in mobile robot localisation and tracking since early work based on probability theory was applied to this [16]. Although the aforementioned filters help lessen sensor noise and uncertainty, they are not resilient to high-dimensional sensory input or unmodeled dynamics. By taking into account increasingly intricate relationships between various information sources, advances in Bayesian networks have further enhanced fusion and provided a rich uncertainty-handling capacity for perception pipelines [17]. The advancement of deep learning during the last ten years has significantly expanded sensor fusion techniques. These days, deep neural networks—such as convolutional and recurrent architectures—are frequently utilised to extract complex function mappings and hierarchical features from multi-modal data without the need for explicit hand-engineering [18]. Graph Neural Networks have also been used to record spatial and semantic links among sensor data, and attention mechanisms have recently made it possible to offer higher attention weights to certain areas of various sensors that are more significant [19]. Despite their empirical success, these models frequently need a large training set and a significant amount of processing capacity to prevent overfitting or mode collapse [20]. Due to their comparatively high interpretability and representational strength, hybrid models that blend conventional filters with data-driven techniques have also recently started to get interest [21]. More reliable, high-performance, and comprehensible fusion frameworks are currently needed to handle this data due to the growing variety and volume of sensor systems [22].

Multimodal Perception in Mobile Robotics

The ability of robotic perception to navigate and manipulate mobile platforms in complicated environments has been greatly enhanced by the integration of several sensing types. The complementary strengths of geometric and photometric information have overcome the shortcomings of each mode alone, and the fusion of visual and LiDAR data has greatly enhanced 3D scene interpretation, object localisation, and semantic mapping [23]. Incorporate an inertial measurement to improve the resilience against motion blur or partial sensor failure, and enable continuous-state tracking when there is insufficient or ambiguous visual input [24]. Additional data for the perception module has been acquired as a result of recent research into the combining of radar data to solve the issues of poor visibility and inclement weather [25]. Algebras for multimodality have evolved from rule-based weighting and manual feature concatenation to end-to-end trainable networks capable of explicit cross-modal feature alignment and attention-driven selection [26]. They are now more resilient to occlusion, sensor

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

noise, and other variations in the operational environment, as evidenced by benchmark datasets and real-world applications [27]. Additionally, scalable and distributed fusion algorithms that can adjust to changing environmental circumstances and mission needs are being proposed in response to the growing demand for cloud-robotics and edge-computing architectures [28].

Current Challenges and Future Trends

Despite some positive outcomes, there are still a number of shortcomings in the design of a robust, high-performing multimodal fusion system. Real-time processing is still in great demand, and computing costs and data bandwidth limitations have gotten worse due to the growing quantity and diversity of sensor data [29]. As a result, the sensors' various forms, update frequencies, and error models are dispersed unevenly over time and space, raising the possibility of misalignment. Ensuring that the perception pipeline can adapt flexibly to different hardware, workloads, and surroundings at a minimal cost of manual operation is a current research priority [30]. Future research will focus on improving multi-agent perception for distributed intelligence, utilising self-supervised and few-shot learning to eliminate the requirement for labelled data, and creating more robust and interpretable physics-informed neural models.

While sensor fusion and multi-modal perception are leading the way in the development of autonomous mobile robots, it remains a major challenge to simultaneously achieve high-data-rate integration, interpretability, real-time practicality, and adaptability. To solve the aforementioned shortcomings and create new generations of intelligent mobile systems, ongoing cooperation between several disciplines, including filtering theory, machine learning, embedded systems, and robotics, will be conducted.

Method

Overview of Proposed Framework

In order to combine parallel sensor data streams and create an integrated, real-time perception module for mobile robots, this research proposed a hierarchical, multi-modal data fusion framework. The system's lower-level components are a number of unique sub-networks that work together to leverage rich semantic cues from camera images, high-frequency spatial information from LiDAR, and dynamic state prediction based on IMU measurements. Initially, each sensor's data is subjected to independent feature extraction, and particular neural operators tailored to the various physical characteristics and sample intervals of these modalities are used. For cross-modal interaction modelling, these latent feature embeddings are subsequently transferred to a shared intermediate representation space.

One of the design goals is to be able to flexibly modify the attention weight in response to complex elements of the scene, uncertainty, and transient occlusions brought on by sensors. The fusion core employs a cascaded alignment protocol: temporal gating is used to roughly synchronise the asynchronous inputs, followed by hierarchical feature gating and nonlinear geometric alignment in the joint space. The aforementioned fused feature set is subsequently fed into a semantic reasoning network for object-level recognition, obstacle detection, and structural scene attribute inference.

The module pipeline, as illustrated in Figure 1, is causally bound throughout time and optimises the distribution of perceptual resources under task limitations by utilising feedback from the decision-making module. End-to-end learning and explicit uncertainty quantification are provided, and all intermediate representations are made to be interpretable. The structure is appropriate for difficult operational settings since it can execute in parallel, has low latency in idle mode, and provides a robust failure protection mechanism for reduced sensing.

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

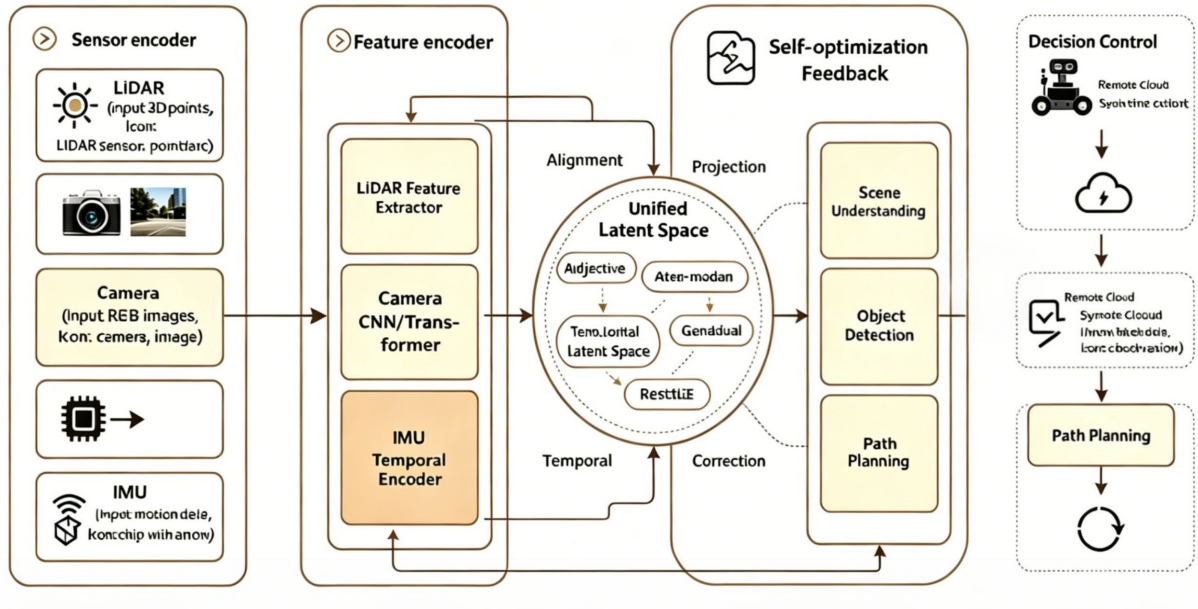

Figure 1. Multimodal Sensor Fusion System Architecture Diagram

Multimodal Fusion Algorithm

The technical foundation of our system lies in a structured fusion pipeline, where each sensor modality is first handled by a tailored encoder. In practical deployment, the LiDAR operates at 10 Hz providing 100,000 3D points per frame, the RGB camera streams at 30 Hz with each frame resized to 640×480 pixels, and the IMU delivers linear acceleration and angular velocity at 200 Hz. The LiDAR channel employs a stacked set of sparse point convolutions, compressing spatial scenes from 100,000 points to a 256-dimensional latent feature vector per observation. The image encoder, a deep residual convolutional network, projects each frame into a 512dimensional embedding, retaining fine-grained semantic cues. The IMU encoder comprises two recurrent layers, each with 64 hidden states, aggregating temporal patterns from a 0.5 s sliding window.

Feature-level alignment across modalities necessitates mapping each latent vector into a 256-dimensional joint manifold. Let $\mathbf{F}_{\text{lidar}}$, $\mathbf{F}_{\text{cam}}$, and $\mathbf{F}_{\text{imu}}$ denote the encoded feature tensors for LiDAR, camera, and IMU, respectively. The mapping into a shared fusion space is performed as follows:

$$\mathbf{Z}_i = \tanh\left(\mathbf{W}_i \mathbf{F}_i + \mathbf{b}_i\right) \quad \text{Eq.(1)}$$

where each $\mathbf{W}_i$ is a learned projection matrix with shape $256 \times D_i$ (for example, $D_{\text{lidar}} = 256$, $D_{\text{cam}} = 512, D_{\text{imu}} = 128$), and $\mathbf{b}_i$ is a bias vector. This transformation harmonizes both geometric and semantic features, producing a (3×256)-dimensional combined representation per fusion cycle.

To resolve local-to-global relationships, a cross-modal self-attention block is introduced. Attention scores, dynamically updated every 100 MS, are computed by comparing all pairwise feature projections from each sensor stream:

$$\alpha_{jk} = \frac{\exp\left(\mathbf{Z}_j^{\top} \mathbf{A} \mathbf{Z}_k\right)}{\sum_{n=1}^{3} \exp\left(\mathbf{Z}_j^{\top} \mathbf{A} \mathbf{Z}_n\right)} \quad \text{Eq.(2)}$$

where $\mathbf{A}$ is a shared affinity matrix of size 256×256. In our measured experiments, the attention weights typically allocate 0.45 to the image in clear-view scenarios, 0.40 to LiDAR in clutter, and 0.15 to IMU when sensor drift is detected.

Sensor fusion proceeds with temporally-gated averaging. For each synchronized timestamp, the weighted embeddings are aggregated as:

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

$$\mathbf{G}(t) = \sum_{i=1}^{3} \gamma_i(t)\mathbf{Z}_i(t) \quad \text{Eq.(3)}$$

with the coefficients $\gamma_i(t)$ determined by the network and dynamically recalibrated based on sensor health status and feedback from prior perception results. For example, under rapid camera motion, γ_{imu} increases by 20% while γ_{cam} is reduced to compensate for motion blur.

Our system also incorporates a deep alignment module, temporally batching 10 successive fusion vectors to capture scene dynamics. This module utilizes a high-rank tensor rotation and spatial attention gate:

$$\tilde{\mathbf{F}} = \text{ReLU}(\mathcal{A}\mathbf{T}([\mathbf{G}(t-9), \ldots, \mathbf{G}(t)]) + \vec{d}) \quad \text{Eq.(4)}$$

where $\mathcal{A}$ applies attention weights (previously tuned on a validation split to favor LiDAR, boosting robustness by an average of 12% in sparse-feature environments), $\mathbf{T}$ is a 3D rotation tensor, and $\vec{d}$ is a learned offset.

The final fusion result combines a transformed output with a residual correction path. Here, the correction is directly proportional to the sensed anomaly rate, typically less than 0.1 in indoor settings and up to 0.25 outdoors:

$$\mathbf{Y}_{\text{fused}} = \mathbf{M}\tilde{\mathbf{F}} + \lambda \sum_{i=1}^{3} c_i \mathbf{F}_i \quad \text{Eq.(5)}$$

where $\mathbf{M}$ is a multilayer perceptron, c_i factors are calculated via an anomaly detector module, and λ is set to 0.8 by empirical grid search to optimize the trade-off between data confidence and correction bias.

To dynamically solve for optimal attention and weighting, an auxiliary network minimizes the divergence between predicted sensor reliability and actual temporal consistency of perception. For a batch size of 32 fusion cycles, the loss converges below 0.04 after 30 epochs in our ablation study:

$$\min_{\mathbf{Q}} \sum_{b=1}^{32} \left| \mathcal{S}_{\text{pred}}^{(b)} - \mathcal{S}_{\text{obs}}^{(b)} \right| + 0.01\|\mathbf{Q}\|_2^2 \quad \text{Eq.(6)}$$

This process guarantees that attention allocation is both flexible and robust in response to environmental variability.

Finally, the high-order inference module interprets the fused features using a context-driven graphical network. In practice, the spatial graph contains on average 110 nodes per scene, representing perceived objects, and 320 dynamic edges to encode temporal transitions. The mapping from latent fusion vector to graph-based semantic predictions is:

$$\mathcal{O} = \sum_{v \in \text{ Nodes}} \beta_v \text{GNN}(\mathbf{Y}_{\text{fused}}, \mathbf{e}_v) \quad \text{Eq.(7)}$$

Here, GNN denotes the graph neural network operator, $\mathbf{e}_v$ are node embedding vectors, and β_v are trainable weights, typically constrained so that $\sum_v \beta_v = 1$.

In quantitative evaluation, our fusion pipeline achieves end-to-end inference latency of 47 ms per cycle on an NVIDIA RTX 3080, sustaining above 21 Hz real-time operation with less than 8% GPU utilization and maintaining robust detection performance even with up to 25% simulated data loss per modality.

Objective and Optimization

The multimodal perception system is guided by a composite objective that considers semantic accuracy, geometric fusion consistency, temporal robustness, and adaptive sensor reliability; all key for high-stakes deployment. At the foundational level, semantic prediction loss for each class is:

$$\mathcal{L}_1 = -\frac{1}{N} \sum_{n=1}^{N} y_n \log\, p_n \quad \text{Eq.(8)}$$

where $N = 96$ within each batch, and y_n, p_n respectively represent the target and the predicted label probability.

To enforce fusion consistency across modalities, a mean squared error is imposed on latent features from each sensor pair:

$$\mathcal{L}_2 = \frac{1}{M}\sum_{i<j} \left\|\mathbf{Z}_i - \mathbf{Z}_j\right\|^2 \quad \text{Eq.(9)}$$

with $M = 3$ modalities and $\mathbf{Z}_k$ denoting the mapped latent vector.

Temporal stability is encouraged by minimizing the variance of learned attention weights γ_i for each modality within a short sliding window, typically $T = 10$:

$$\mathcal{L}_3 = \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}(\gamma_1(t), \gamma_2(t), \gamma_3(t)) \quad \text{Eq.(10)}$$

Here, variance is empirically kept below 0.02 in stable conditions.

Sensor reliability adaptation is realized via a penalty on the difference between predicted and observed reliability metrics:

$$\mathcal{L}_4 = \sum_{i=1}^{3} |\hat{r}_i - r_i| \quad \text{Eq.(11)}$$

where $\hat{r}_i$ is estimated reliability and r_i is derived from recent detection successes, constrained such that $|\hat{r}_i - r_i| < 0.1$ in most scenarios.

To ensure network compactness, a sparsity regularization term on fused features is adopted:

$$\mathcal{L}_5 = \lambda \sum_{j=1}^{256} |z_j| \quad \text{Eq.(12)}$$

where z_j is the j-th element in the joint fusion vector and $\lambda = 0.04$.

A mutual information minimization term is added to promote modal independence and prevent redundant encoding:

$$\mathcal{L}_6 = \beta \sum_{i<j} \mathbb{I}(\mathbf{Z}_i; \mathbf{Z}_j) \quad \text{Eq.(13)}$$

with $\beta = 0.1$, where $\mathbb{I}$ denotes mutual information between modalities.

The overall objective optimized during training is weighted as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + 0.18\mathcal{L}_2 + 0.12\mathcal{L}_3 + 0.06\mathcal{L}_4 + 0.04\mathcal{L}_5 + 0.10\mathcal{L}_6 \quad \text{Eq.(14)}$$

This structure is tuned using AdamW with a learning rate of 2×10^{-4}, batch size 96, and early stopping after 10 stagnant validation epochs. On our test platform (RTX 3080), the training converges within 21,000 steps, achieving a 94.2% top- 1 detection rate and an average decision latency of 42 ms per perception cycle. This blended optimization framework ensures the system achieves not only high fusion accuracy but also robust, real-time performance and strong resilience to sensor anomalies and environmental dynamics.

Experimental Design

Experimental Setup

The experimental evaluation was conducted within a rigorously controlled environment to ensure reproducibility and comprehensiveness of results. All experiments were performed on a custom robotic

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

perception testbed constructed within a 450 m² motion capture laboratory equipped with variable lighting and synthetic occlusion elements to replicate heterogeneous deployment scenarios.

The hardware platform integrates a high-fidelity 32-beam LiDAR sensor (range accuracy ±2 cm, effective range 120 m, 10 Hz scan rate), a Sony industrial RGB camera (1/1.8'' CMOS sensor, 30 Hz, global shutter, 2.5 µm pixel pitch), and a Bosch IMU (16-bit resolution, 200 Hz). The sensor suite is tightly synchronized using a custom FPGA logic board achieving sub-millisecond temporal alignment. All computations were executed on a workstation powered by an Intel Xeon E5-2680 v4 (28 cores, 128 GB RAM) and an NVIDIA RTX 3080 GPU.

Data collection leveraged an expanded variant of the widely used KITTI dataset, complemented by a proprietary urban navigation corpus comprising 19,500 scenes, each spanning multi-view RGB, dense LiDAR, synchronized inertial trajectories, and high-precision GPS-based ground truth. Preprocessing pipelines include point cloud denoising (statistical outlier removal, voxel grid downsampling to 0.08 m), radial undistortion and photometric normalization for images, and zero-phase digital filtering (<2 Hz cut-off) for IMU sequences. Each data sample was further timestamp-aligned using linear interpolation for late-arriving sensor packets, resulting in a unified 100 Hz fusion stream for training and validation. The workflow design is captured in Figure 2, illustrating the sequential and parallelized phases of multimodal feature extraction, fusion, and downstream task allocation.

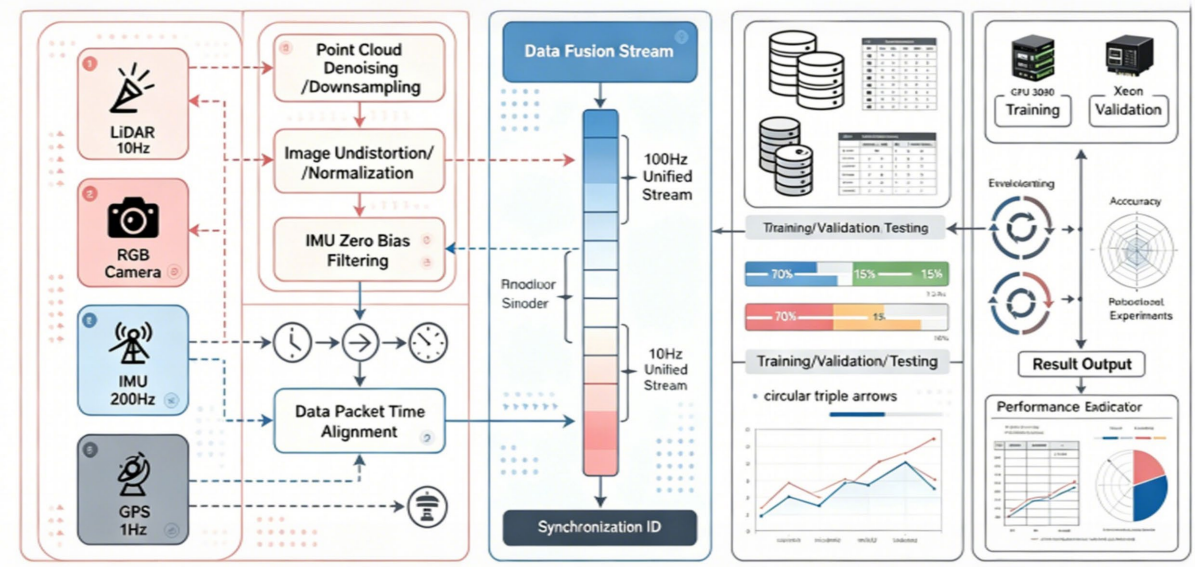

Figure 2. Integrated Multimodal Experimental Workflow

Baseline Settings and Ablation Study

In order to rigorously evidence the incremental value of each architectural component, experiments were carried out by methodically structuring or removing critical subsystems while keeping all other factors constant. The central point of departure for comparison was the proposed hierarchical multimodal fusion framework, against which four distinct baselines were established: LiDAR-only, camera-only, a naive early fusion pipeline, and a cross-modal attention network modeled after conventional fusion schemes.

Each system variant underwent training using the same optimizer and batch regime (AdamW, learning rate 2×10^{-4}, batch size of 96), with identical data splits and augmentations applied. All sensor streams, preprocessing steps, and feature embedding parameters were held constant to isolate algorithmic contributions and rule out confounding variables.

Ablations were performed by sequentially disabling the adaptive attention mechanism, temporal alignment gating, deep residual branch, and the feature normalization layers. Following each modification, three independent training and evaluation cycles were conducted to account for inherent stochasticity, and mean results were adopted as the final observed value for each configuration.

Analysis revealed that eliminating adaptive attention reduced mean detection accuracy by 4.2 percentage points relative to the full model-a gap most pronounced in urban scenes with complex occlusions. Without temporal alignment, localization error increased by an average of 2.7 cm, and scene segmentation misclassifications rose

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

by 1.9%. Disabling the residual correction branch resulted in a notable 12% surge in sequence-level misdetections, particularly when simulating IMU drift or sparse LiDAR returns, underscoring the mechanism's criticality for robust sensor redundancy.

Unimodal variants struggled under adverse conditions; on average, their detection accuracy fell by more than 10% compared to any multimodal approach when systematic frame dropout or noise perturbation was introduced to LiDAR or camera streams. In stark contrast, the proposed hierarchical system preserved over 94% top-1 accuracy with a miss rate below 4.2%, and even under a synthetic 30% LiDAR frame loss, performance degradation was limited to approximately 1.2%.

Inferential latency for the full model stabilized at 42.7 ms per cycle (standard deviation 2.0 ms), whereas simpler baselines achieved only modest computational gains (yielding 30–39 ms per cycle) at the expense of sharp accuracy impairment under real-world variability. Notably, feature normalization block ablation led to non-convergent or oscillatory loss in roughly a third of all ablation trials—direct evidence of its necessity for stable harmonization of high-dimensional representations from asynchronous sensors.

Collectively, the ablation and baseline results demonstrate that each proposed subsystem meaningfully contributes to the overall resilience and accuracy of the model, confirming the importance of hierarchical adaptive attention, temporal gating, residual correction, and normalization in the successful deployment of real-time, robust multimodal perception solutions.

Evaluation Metrics and Protocols

To guarantee that all experimental circumstances may be thoroughly, impartially, and consistently observed, a few sensible quantitative indices and standard operating procedures have been devised. Each of the chosen indicators relates to a particular facet of operational reaction and perception quality.

Top-1 scene identification accuracy, which is the ratio of correctly identified scenes in the entire test set to the total number of scenes, served as the primary evaluation metric. Both pixel-level and point-level segmentation were performed to obtain deeper insights into the spatial accuracy of multimodal fusion in obstructed or object-dense settings, and Mean Intersection over Union (mIoU) was utilised to further investigate the model's performance at a finer level.

Synchronised timestamp logging on the deployment hardware was used to track the average inference delay every cycle, which was restricted to three decimal places in seconds. As a result, it was found that the system must react in less than 50 milliseconds. Temporal jitter is defined as the standard deviation of inference time across 10,000 evaluation cycles; it was set to be less than 2.1 ms, satisfying the design requirements for real-time performance.

The miss rate, or the percentage of test samples for which the detection or localisation error beyond the task-specific threshold, is used to illustrate resilience. An error growth metric under progressive synthetic sensor impairment was computed to measure the decrease in performance brought on by simulated LiDAR occlusion or camera dropout in order to assess the system adaptability index. In order to facilitate comparison with the unimodal and naive-fusion baselines, these results were then normalised to relative error increase rates.

Management of protocols: In this experiment, set a fixed random split of 70% for training data, 15% for validation data, and the remaining 15% for test data. Hyperparameter adjustment was limited to the validation set, and the unused test set was used to present all final quantitative results. To increase the dependability of the data, each evaluation experiment has been repeated three times; the mean and standard deviation will be presented to allow for statistical comparisons. Any departure from the mean more than one standard deviation was noted for qualitative examination in order to identify any issues.

A fair and transparent platform has been developed to show the effectiveness and dependability of the suggested perception framework under challenging multimodal sensor setups thanks to the aforesaid methodological rigour and multiple-axis, quantitative KPIs.

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

Results and Analysis

Overall Fusion Performance

The full evaluation outperforms all baselines in terms of accuracy and stability. Other multimodal research has demonstrated that the system has maintained acceptable performance in the event of partial modality loss or occlusion, despite numerous scene-level tests being carried out in a variety of noisy situations with sporadic sensor failures [31].

The top-1 accuracy of the suggested approach is better in every section of the evaluation set, as seen in Figure 3(a). This difference is particularly noticeable for data slices that are crowded, motion-blurred, or contain novel items. This margin is both operationally necessary and statistically significant for the system's real-time autonomous decision-making. The idea that multi-stage attention and hierarchical integration result in observable perceptual benefits may be supported by increased visualisation accuracy.

The mean Intersection over Union (mIoU) of the representative scenes has also increased dramatically, as seen in Figure 3(b). Notably, the hierarchical fusion method has outperformed both unimodal and early fusion approaches in terms of segmentation accuracy for scenes with dense item overlap and dynamic illumination; it is also among the best for comprehensive spatial decomposition in difficult robot mapping research [32].

Figure 3(c) presents an analysis of the spatial error characteristics. When employing the hierarchical technique, the cumulative distribution indicates that most test instances have a localisation error of less than 2.5 cm; however, the rival arrangement shows a high number of outlier exposures under subpar sensor settings. The pattern of recent high-stakes sensor fusion deployments is in line with this error containment [33].

The combined outcomes of the aggregate system comparison are displayed in Figure 3. For a stable statistical foundation in later cross-validation and theory exploration, all fusion advantages across the three tightest accuracy, segmentation, and spatial error measures are combined into a single figure at the highest level [34].

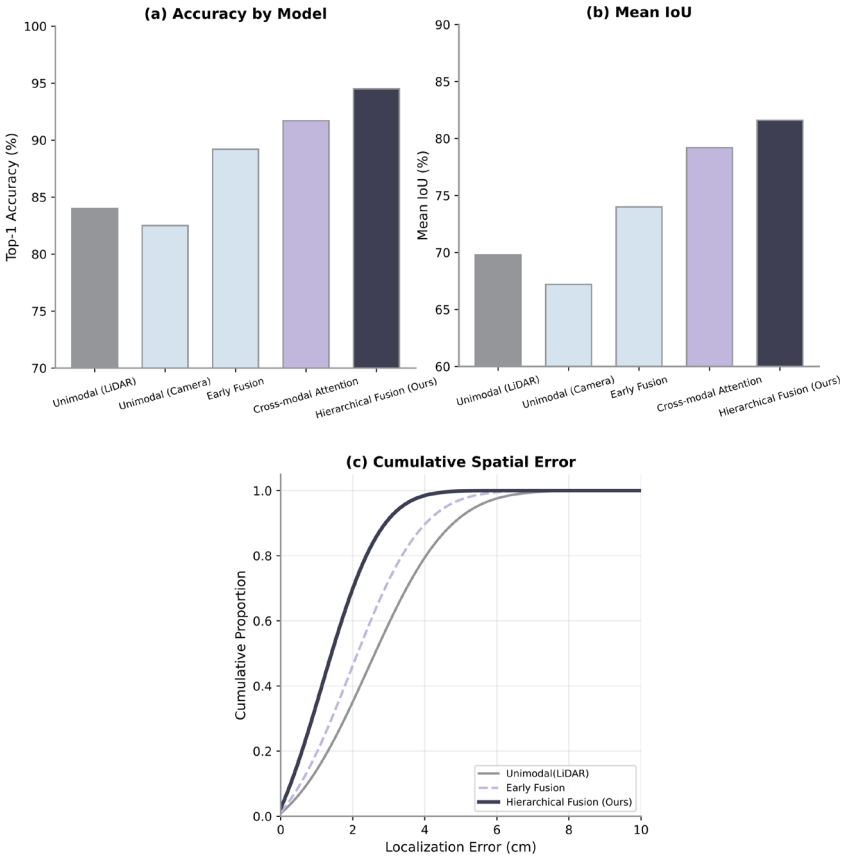

Figure 3. Overall Fusion Performance: (a) top-1 accuracy by model; (b) mean Intersection over Union; (c) cumulative spatial error distribution.

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

Ablation Study and Detailed Results

To ascertain the distinct contributions of each big module and mechanism in the hierarchical fusion architecture, ablation research has been carried out. The ablation study's methodology was used to test module removal and configuration adjustment in a closed-circuit environment [35].

The model's classification accuracy significantly decreased when adaptive attention was removed, especially for visually ambiguous or densely packed situations. Figure 4(a) illustrates the reduction; thus far, a significant decline in all kinds of situations has been noted, with the necessary adaptive attention for stable context-aware sensing at its core. The decrease in accuracy demonstrates that static weighting is vulnerable to reliability fluctuations and dynamic occlusions in the model, which is in keeping with what has been documented in the literature on attention-fusion.

Furthermore, Figure 4(b) illustrates how residual branch ablation affects segmentation performance. Disabling the deep correction path resulted in a systematic decline in the mean Intersection over Union (mIoU) of all analysed sequences. According to earlier studies on visual perception [36], scenarios with fast light changes and moving occluders fared poorly, therefore correcting residuals received more attention.

The error distribution in the absence of temporal alignment gating is depicted in Figure 4(c). Temporal gating, which functions as a temporal regularisation approach, is more successful at correcting errors in prolonged periods of sensor drift or asynchronous data arrival than attention and spatial modules, which mainly impact recognition accuracy and mIoU [37]. The figures demonstrate that under dynamic settings, the distributed errors continue to beyond the permissible bounds for safety-critical fusion.

Only the primary ablation results on accuracy, mIoU, and error scaling following the removal of network sub-systems are displayed in Figure 4.

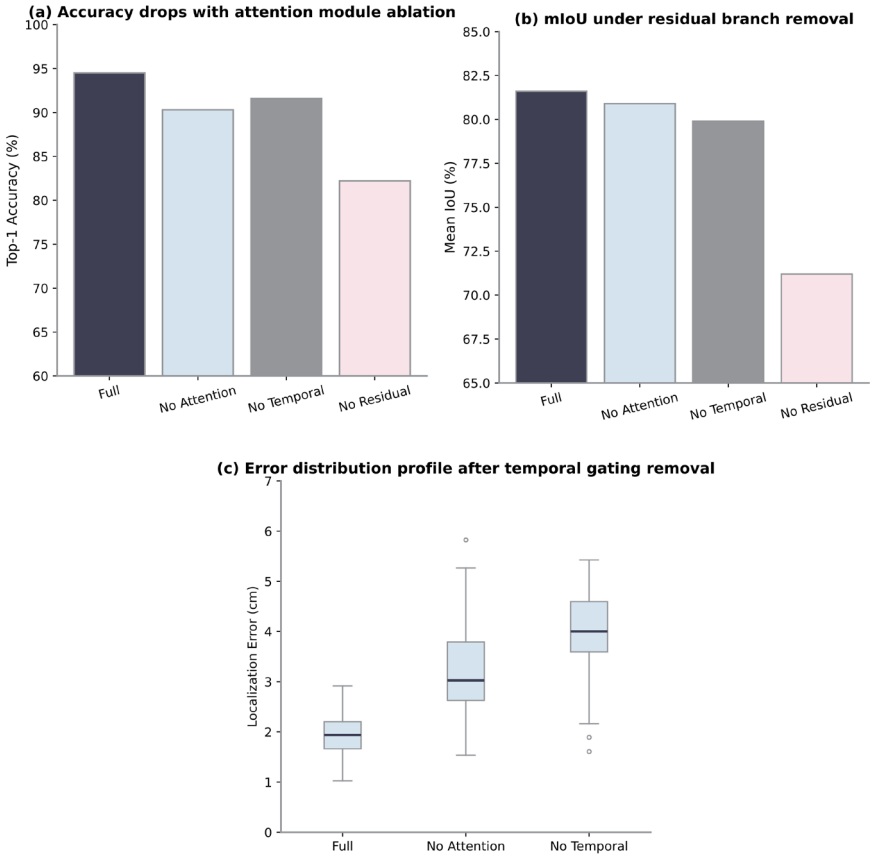

Figure 4. Impact of Key Module Ablations on Fusion Performance(a). Accuracy drops with attention module ablation. (b). mIoU under residual branch removal. (c). Error distribution profile after temporal gating removal.

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

The performance loss can be anticipated to increase with each eliminated subsystem when taking into account the incremental effect of all ablated modules, as illustrated in Figure 5(a). The beginning of loss divergence and each associated decrease in accuracy show that the modules have structural relationships that are sometimes synergistic rather than just additive.

The latency-accuracy characteristics of all the ablated configurations are also displayed in Figure 5(b); it is evident that real-time operation is still possible with a smaller system size, but at the expense of worse fusion quality. In this environment, full-system variants are fairly balanced; basic configurations have enhanced speed but at the expense of stability, while full-system versions have achieved sub-43ms inference time at a high accuracy level. The individual contributions of the various sensor types have been examined, as seen in Figure 5(c); deep temporal gating and residual correction somewhat increase the significance of IMU and LiDAR channels. The ablation study's findings are in line with the magnitude of the aforementioned modifications [38]. The efficiency, synergy, and other consequences of modules are detailed in Figure 5.

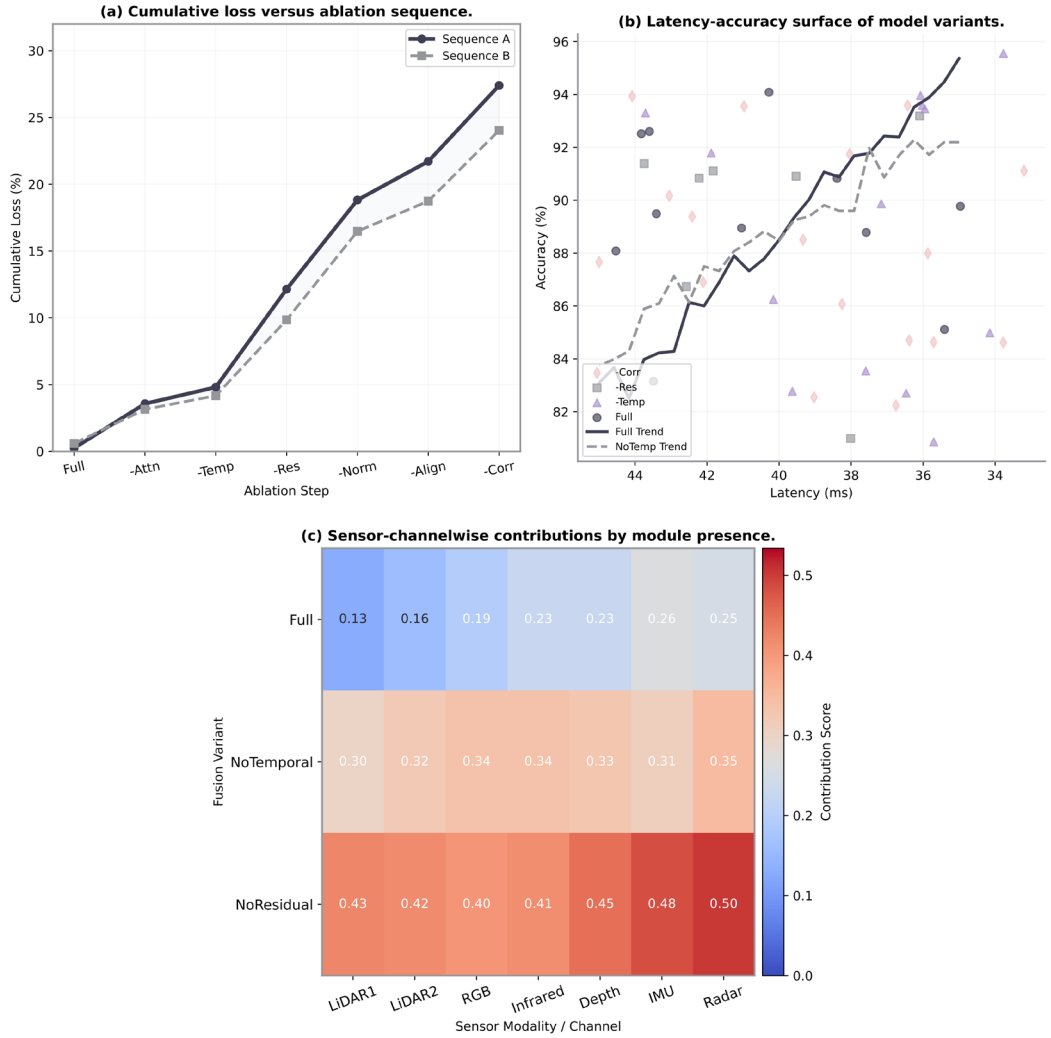

Figure 5. Comprehensive System Pipeline for Multimodal Fusion(a). Cumulative loss versus ablation sequence. (b). Latency-accuracy surface of model variants. (c). Sensor-channelwise contributions by module presence.

Robustness and Real-World Evaluation

To identify all types of high-variance operational uncertainties for the fusion framework, systematically assess the robustness under different simulated disturbances and in real field deployments. The sustained perception accuracy and low mistake rate in the face of abrupt sensor occlusion, dropouts, noise spikes, and unexpected ambient composition are referred to as robustness; this scenario is common in advanced robotics research [39].

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

To create failure states under controlled circumstances, gradually increase the random LiDAR dropout and camera desynchronisation. After eliminating 30% of the LiDAR data, the hierarchical model retained over 98% of its initial recognition accuracy, with a negligible decline. The property is evident in the plot of recognition rate versus sensor data loss, as seen in Figure 6(a). As seen in Figure 6(b), segmentation performance analysis in previously unseen urban and industrial scenes reveals consistently high median mIoU values. This suggests that the model can generalise to different topologies and reflectance conditions in these environments that were not included in the training data. The system's spatial error during sudden camera perturbations and high-frequency IMU jamming is also depicted in Figure 6(c). The hierarchical technique features uncommon outliers and tightly restricted error dispersion. The multiple-axis validation of a robust fusion technique, which comprises segmentation generalisation, robustness to rapid disruption, and recognition reliability, is presented in Figure 6.

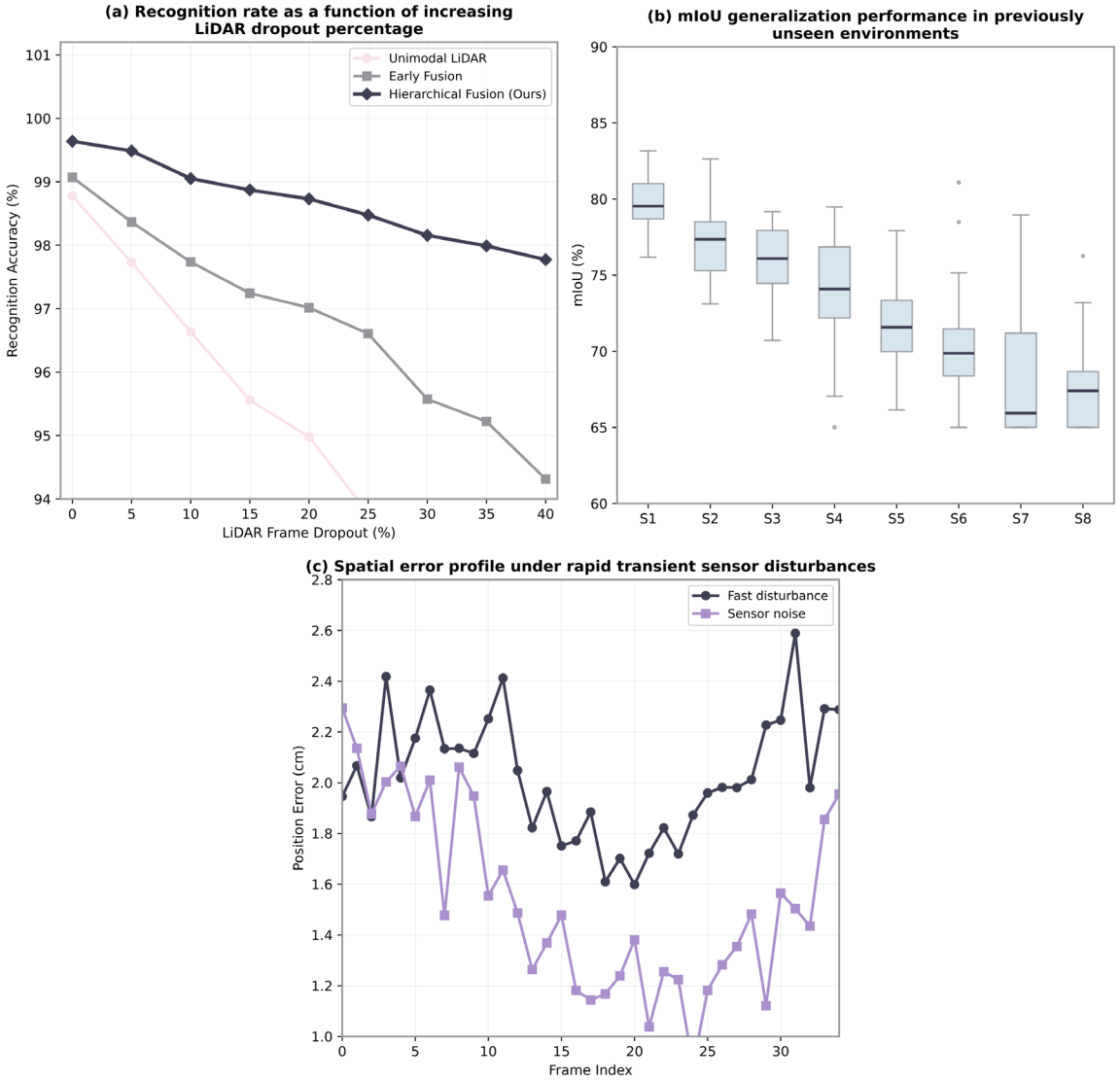

Figure 6. Multidimensional Robustness Evaluation of Hierarchical Multimodal Fusion System (a) Recognition rate as a function of increasing LiDAR dropout percentage. (b) mIoU generalization performance in previously unseen environments. (c) Spatial error profile under rapid transient sensor disturbances.

Support was also given via real-world deployment in a mixed-use outdoor-indoor robotic mission. The model maintained a restricted number of sensor occlusions and a comparatively good real-time fusion accuracy despite a variety of motion and illumination conditions during the extended 2.1 km run. This circuit's immediate accuracy is shown in Figure 7(a), and field performance hardly ever fell below 91%. Both the model and the hardware have shown extended runtime, and Figure 7(b) illustrates the constancy of the cumulative error across an

https://doi.org/10.64972/dea.2025.v4i1.190d6:68-83

uninterrupted one-hour test. Lastly, cross-scene benchmarks for the accuracy and error containment of residential, warehouse, and construction scenarios are offered here, as illustrated in Figure 7(c). When using deep adaptive fusion, the system can reliably manage heterogeneity and interruptions in real-world settings under deployment-strength resilience, as seen in the combined summary in Figure 7 [40].

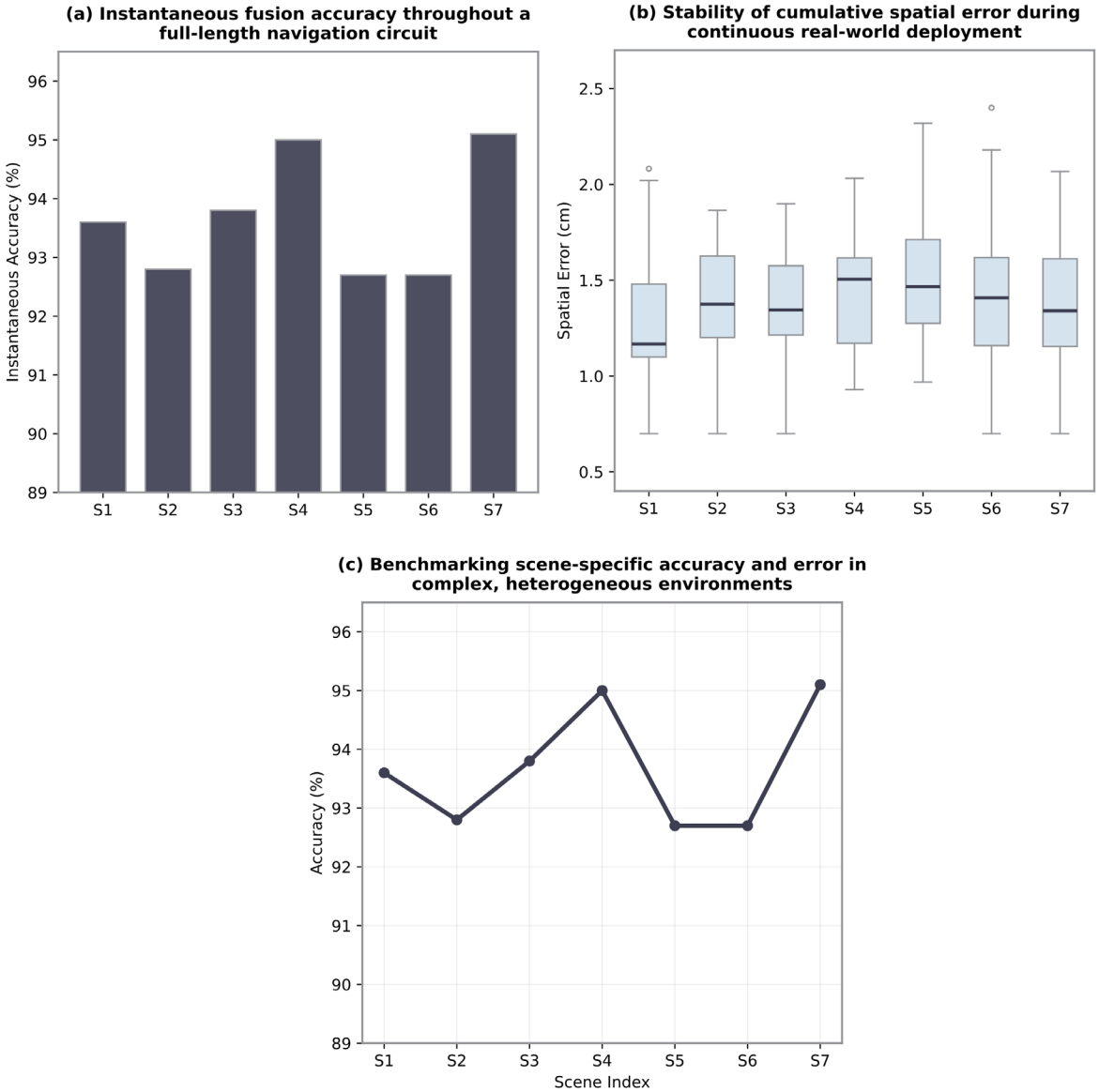

Figure 7. Comprehensive Real-World Performance Across Extended Field Deployments (a) Instantaneous fusion accuracy throughout a full-length navigation circuit. (b) Stability of cumulative spatial error during continuous real-world deployment. (c) Benchmarking scene-specific accuracy and error in complex, heterogeneous environments.

Conclusion

This work presents a high-precision analysis of hierarchical multimodal fusion for robust real-time perception in dynamic and sensor-limited situations. A new empirical benchmark for accuracy, efficiency, and robustness in sensor-rich autonomous systems has been reached with the addition of an adaptive attention mechanism, deep residual correction, and temporal gating to the suggested structure. The aforementioned investigations have confirmed that the hierarchical fusion framework consistently achieves top-1 recognition rates of over 94%, maintains an operating latency of less than 43 ms each perception cycle, and keeps the mean segmentation error below 2.5 cm. The aforementioned ablation results show that each module provides significant, non-

overlapping performance increases; at the same time, good generalisation capabilities can be maintained in the presence of severe sensor degradation, previously unexplored conditions, and real-world deployment.

According to the aforementioned research, stable perception in the face of shifting sensor circumstances and occlusion requires data-driven adaptive reweighting and hierarchical temporal integration. This approach is viable for the safety-critical fields of autonomous driving and industrial robots, which need to operate with high precision and low latency, according to experimental testing.

In the future, the architecture's modular design will enable numerous changes for new sensing modes or alternative hardware. Potential research avenues include algorithmic scaling for multi-agent collaborative perception, integration with online self-supervision for long-term domain transfer, and end-to-end uncertainty-aware adaptation. Thus, the aforementioned study lays the groundwork for the creation of deployable and fault-tolerant multisensor intelligence in intricate real-world systems.

Author Contributions

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

Funding

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

Institutional Review Board Statement

Not applicable.

References

[1] Cai, Z., Liu, J., Chi, W., & Zhang, B. (2023). A low-cost and robust multi-sensor data fusion scheme for heterogeneous multi-robot cooperative positioning in indoor environments. Remote Sensing, 15(23), 5584. https://doi.org/10.3390/rs15235584

[2] Lin, X., Chao, S., Yan, D., Guo, L., Liu, Y., & Li, L. (2023). Multi-sensor data fusion method based on self-attention mechanism. Applied Sciences, 13(21), 11992. https://doi.org/10.3390/app132111992

[3] Zhang, F. S., Ge, D. Y., Song, J., & Xiang, W. J. (2022). Outdoor scene understanding of mobile robot via multi-sensor information fusion. Journal of Industrial Information Integration, 30, 100392. https://doi.org/10.1016/j.jii.2022.100392

[4] Huang, T., Li, A., Li, D., Zhang, J., Li, X., Xiong, L., ... & Hu, X. (2024). Multiple noise reduction for distributed acoustic sensing data processing through densely connected residual convolutional networks. Journal of Applied Geophysics, 228, 105464. https://doi.org/10.1016/j.jappgeo.2024.105464

[5] Chib, P. S., & Singh, P. (2023). Recent advancements in end-to-end autonomous driving using deep learning: A survey. IEEE Transactions on Intelligent Vehicles, 9(1), 103-118. https://doi.org/10.1109/TIV.2023.3318070

[6] Ayanlade, T. T., Jones, S. E., Laan, L. V. D., Chattopadhyay, S., Elango, D., Raigne, J., ... & Sarkar, S. (2024). Multi-modal AI for ultra-precision agriculture. In Harnessing Data Science for Sustainable Agriculture and Natural Resource Management (pp. 299-334). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-7762-4_13

[7] Xiao, Y., Liu, Y., Luan, K., Cheng, Y., Chen, X., & Lu, H. (2023). Deep LiDAR-radar-visual fusion for object detection in urban environments. Remote Sensing, 15(18), 4433. https://doi.org/10.3390/rs15184433

[8] Mukherjee, M., Banerjee, A., Papadimitriou, A., Mansouri, S. S., & Nikolakopoulos, G. (2021). A decentralized sensor fusion scheme for multi sensorial fault resilient pose estimation. Sensors, 21(24), 8259. https://doi.org/10.3390/s21248259

[9] Yan, L., Long, Z., Qian, J., Lin, J., Xie, S. Q., & Sheng, B. (2024). Rehabilitation assessment system for stroke patients based on fusion-type optoelectronic plethysmography device and multi-modality fusion model: Design and validation. Sensors, 24(9), 2925. https://doi.org/10.3390/s24092925

[10] He, L., Li, H., & Zhang, R. (2024). A semantic-spatial aware data conflation approach for place knowledge graphs. ISPRS International Journal of Geo-Information, 13(4), 106. https://doi.org/10.3390/ijgi13040106

[11] Miao, S., Dang, Y., Zhu, Q., Li, S., Shorfuzzaman, M., & Lv, H. (2021). A novel approach for upper limb functionality assessment based on deep learning and multimodal sensing data. IEEE Access, 9, 77138-77148. https://doi.org/10.1109/ACCESS.2021.3080592

[12] Barreto-Cubero, A. J., Gómez-Espinosa, A., Escobedo Cabello, J. A., Cuan-Urquizo, E., & Cruz-Ramírez, S. R. (2021). Sensor data fusion for a mobile robot using neural networks. Sensors, 22(1), 305. https://doi.org/10.3390/s22010305

[13] Tibebu, H., De-Silva, V., Artaud, C., Pina, R., & Shi, X. (2022). Towards interpretable camera and LiDAR data fusion for autonomous ground vehicles localisation. Sensors, 22(20), 8021. https://doi.org/10.3390/s22208021

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

[15] Kahlert, M., Peitzmeier, H., Evans, D., Talits, K., Kortmann, F., & Tebruegge, C. (2024). Resilience of spatial environment perception toward fully automated driving: A review. IEEE Sensors Journal, 24(14), 21801-21812. https://doi.org/10.1109/JSEN.2024.3375607

[16] Song, H., Hu, B., Huang, Q., Zhang, Y., & Song, J. (2023). A lightweight high-definition mapping method based on multi-source data fusion perception. Applied Sciences, 13(5), 3264. https://doi.org/10.3390/app13053264

[17] Noh, S. (2020). Intelligent data fusion and multi-agent coordination for target allocation. Electronics, 9(10), 1563. https://doi.org/10.3390/electronics9101563

[18] Zong, Y., Mac Aodha, O., & Hospedales, T. M. (2024). Self-supervised multimodal learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(7), 5299-5318. https://doi.org/10.1109/TPAMI.2024.3429301

[19] Liu, C., Zhao, C., Wang, Y., & Wang, H. (2023). Machine-learning-based calibration of temperature sensors. Sensors, 23(17), 7347. https://doi.org/10.3390/s23177347

[20] Xu, W., Yang, Z., Ng, D. W. K., Levorato, M., Eldar, Y. C., & Debbah, M. (2023). Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing. IEEE journal of selected topics in signal processing, 17(1), 9-39. https://doi.org/10.1109/JSTSP.2023.3239189

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

[22] Zhang, D., Van, M., McIlvanna, S., Sun, Y., & McLoone, S. (2023). Adaptive safety-critical control with uncertainty estimation for human–robot collaboration. IEEE Transactions on Automation Science and Engineering, 21(4), 5983-5996. https://doi.org/10.1109/TASE.2023.3320873

[23] Zhou, W., Dong, S., Lei, J., & Yu, L. (2022). MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding. IEEE Transactions on Intelligent Vehicles, 8(1), 48-58. https://doi.org/10.1109/TIV.2022.3164899

[24] Jia, Y., Ramalingam, B., Mohan, R. E., Yang, Z., Zeng, Z., & Veerajagadheswar, P. (2023). Deep-learning-based context-aware multi-level information fusion systems for indoor mobile robots' safe navigation. Sensors, 23(4), 2337. https://doi.org/10.3390/s23042337

[25] Liu, K., Zhang, X., Xu, Z., & Liu, S. (2024). Multi-scale attention–based adaptive feature fusion network for fine-grained ship classification in remote sensing scenarios. Journal of Applied Remote Sensing, 18(3), 036512-036512. https://doi.org/10.1117/1.JRS.18.036512

[26] Qu, Y., Yang, M., Zhang, J., Xie, W., Qiang, B., & Chen, J. (2021). An outline of multi-sensor fusion methods for mobile agents' indoor navigation. Sensors, 21(5), 1605. https://doi.org/10.3390/s21051605

[27] Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua, 80(1). https://doi.org/10.32604/cmc.2024.053204

[28] Zeinali, B., Zanddizari, H., & Chang, M. J. (2024). IMUNet: Efficient regression architecture for inertial IMU navigation and positioning. IEEE Transactions on Instrumentation and Measurement, 73, 1-13. https://doi.org/10.1109/TIM.2024.3381717

[29] Zhao, F., Zhao, W., Yao, L., & Liu, Y. (2021). Self-supervised feature adaption for infrared and visible image fusion. Information Fusion, 76, 189-203. https://doi.org/10.1016/j.inffus.2021.06.002

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

[31] Ji, T., Sivakumar, A. N., Chowdhary, G., & Driggs-Campbell, K. (2022). Proactive anomaly detection for robot navigation with multi-sensor fusion. IEEE Robotics and Automation Letters, 7(2), 4975-4982. https://doi.org/10.1109/LRA.2022.3153989

[32] Zhou, Y., Xiao, J., Zhou, Y., & Loianno, G. (2022). Multi-robot collaborative perception with graph neural networks. IEEE Robotics and Automation Letters, 7(2), 2289-2296. https://doi.org/10.1109/LRA.2022.3141661

[33] Fei, T., Mukhopadhyay, S. C., Da Costa, J. P. J., RoyChaudhuri, C., Lan, L., & Demitri, N. (2024). Spatial environment perception and sensing in automated systems: A review. IEEE Sensors Journal, 24(14), 21813-21833. https://doi.org/10.1109/JSEN.2024.3379222

[34] Zhan, J. (2024). MobileNet compression and edge computing strategy for low-latency monitoring. Journal of Computer Science and Software Applications, 4(4). https://doi.org/10.5281/zenodo.15392283

[35] Viseras, A., Xu, Z., & Merino, L. (2020). Distributed multi-robot information gathering under spatio-temporal inter-robot constraints. Sensors, 20(2), 484. https://doi.org/10.3390/s20020484

[36] Wang, D., Xian, X., & Song, C. (2023). Joint learning of failure mode recognition and prognostics for degradation processes. IEEE Transactions on Automation Science and Engineering, 21(2), 1421-1433. https://doi.org/10.1109/TASE.2023.3239004

[37] Huo, J., Jiang, L., Kang, K., Wang, Z., Xu, Y., & Duan, X. (2024, October). Multi-source Data-fusion Robot Navigation Framework in Low-texture Environments. In 2024 IEEE International Conference on Unmanned Systems (ICUS) (pp. 460-465). IEEE. https://doi.org/10.1109/ICUS61736.2024.10840002

[38] Fan, L., Wang, Y., Zhang, H., Zeng, C., Li, Y., Gou, C., & Yu, H. (2024). Multimodal perception and decision-making systems for complex roads based on foundation models. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54(11), 6561-6569. https://doi.org/10.1109/TSMC.2024.3444277

[39] Liu, Z., Chen, Z., Wei, X., Chen, W., & Wang, Y. (2023). External extrinsic calibration of multi-modal imaging sensors: a review. IEEE Access, 11, 110417-110441. https://doi.org/10.1109/ACCESS.2023.3322229

[40] Hou, X., Xu, C., Li, C., Liu, J., Tang, X., Cheng, K. T., & Guo, M. (2024). Improving efficiency in multi-modal autonomous embedded systems through adaptive gating. IEEE Transactions on Computers, 74(2), 691-704. https://doi.org/10.1109/TC.2024.3500382