Enhancing Autonomous Vehicle Perception through Multi-Sensor Fusion and Uncertainty-Aware Decision-Making

Vikash Kumar Verma
M. Tech Scholar
Department of Computer Science
SAM College
Bhopal, Madhya Pradesh, India
jecvikas82@yahoo.com

Dr. Sourabh Mandaloi
Associate Professor
Department of Computer Science
SAM College
Bhopal, Madhya Pradesh, India

***Abstract*:** Autonomous driving requires precise perception and decision making in non-static, complex and uncertain settings. Systems that only rely on one sensor, are typically not reliable in low visibility, occlusion or dynamic conditions, all of which will ultimately affect safe navigation. This research will assess the likelihood that multi-sensor fusion of RGB and depth/LiDAR, can improve the accuracy of perception, in addition to whether a model of uncertainty can be leveraged to heuristically mitigate sensor reliability. Finally, to also improve online decision-making in steering, throttle and braking driving agents. To this end, the proposed Uncertainty-Aware TransFuse fuses RGB and depth features using CNNs with Transformer-based attention. Incorporating uncertainty leverages weighted reliance upon sensor reliability/uncertainty heuristics during inference. Experimental results on the KITTI dataset demonstrated statistically significant improvements over the baseline, in navigation accuracy, object detection performance, and lane detection performance, even when the scenes were severely degraded. The proposed multi-sensor fusion system works in real-time at 32 FPS and reduces false detections in occlusion and low-light testing conditions. In short, multi-sensor fusion of RGB and depth, with uncertainty, increases overall robustness and safety. The Uncertainty-Aware TransFuse provides a robust model of real-world driving for reliable autonomous perceiving.

Keywords: Autonomous Driving, Multi-Sensor Fusion, Uncertainty-Aware Fusion, TransFuse Model, Deep Learning, RGB–LiDAR Fusion

I. INTRODUCTION

As additional robotic systems are being deployed in areas like autonomous driving, industrial automation, healthcare, and service, the demand for robust perception and navigation in a myriad of contexts has increased. Over the years, robot perception has evolved from a rule-based systems that only utilized single sensor modalities (e.g., sonar or camera) to learning-based paradigms that utilize advances in computer vision, probabilistic models, and SLAM [3]. Unfortunately, the relationship between data and perception has changed due to the emergence of deep learning, which has allowed automatic feature extraction and hierarchical representation of sensory data to produce a more robust and adaptive understanding of intricate scenes. Significantly, multi-modal sensor fusion in sensing which leverages uncorrelated/complementary data (LiDAR, RGB camera and an IMU) has become a "go to" approach in overcoming the limitations of single modalities and improving decision making in uncertain and unstructured environments [2][4].

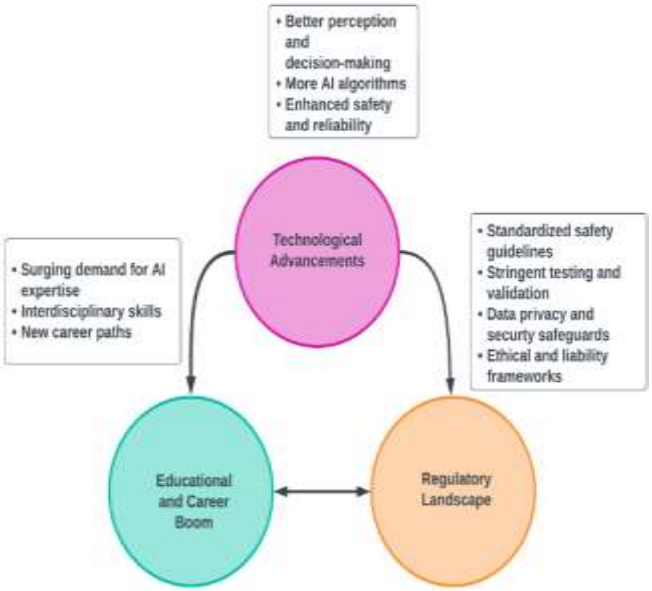

Figure 1. Evolution of Robot Perception [4]

Reliable navigation is critical for autonomous robots executing tasks in, complex and uncertain environments. Navigation entails precise localization, obstacle avoidance, and planning safe motions for paths in the presence of the challenges of sensor noise, dynamic objects, and environmental changes [5]. Single-sensor systems struggle to perform when occluded, or in challenging environments, thereby limiting their reliability in real-world applications. Deep learning-based multi-modal fusion enhances reliability by combining information from complementary sensors such as cameras, LiDAR, and IMUs [6]. Vision–LiDAR fusion increases the robot's understanding of depth information whereas IMUs act as a check on localization and reduce the risk of unreliable velocity estimates thus increasing safety in navigation. This suggests that perception is important as it allows robots to sense, interpret, and either or both physically and informationally interact with the surrounding environment [7] [8]. Sensor-based perception systems (e.g., cameras, LiDAR, sonar, or IMUs) generate awareness of the surroundings in real-time with

Received 18 July 2025, Accepted 23 October 2025

each having their respective advantages and limitations [9] [10]. Using multiple modalities to run perception systems increases the robustness of the navigation and reliability of operation across different and dynamic environments [11].

Table 1. Challenges in Dynamic and Unstructured Environments [12]

Challenge	Description
Sensor Noise and Uncertainty	Environmental factors introduce noise in sensor data, reducing accuracy of perception and navigation.
Occlusion and Clutter	Obstacles and overlapping objects hinder object recognition and scene understanding.
Dynamic Objects	Moving entities such as humans, vehicles, or drones create unpredictable scenarios for navigation.
Environmental Variability	Changing lighting, weather, and terrain conditions degrade the reliability of single-modality sensors.
Real-Time Processing	High-dimensional multi-sensor data requires efficient algorithms to ensure timely decision-making.
Adaptability and Generalization	Difficulty in transferring models trained in one environment to new, unseen, and complex settings.

The integration of different sensing modalities is a crucial part of enhanced robot perceptions as it can fuse different sensor data to produce a more complete understanding of the environment. With complementary sensing, robots can use information from multiple sensors to mitigate the weaknesses inherent with each sensor, thus increasing both the robustness of the robot's perception and improved decision-making under dynamic uncertain states [13]. Single-modality systems have inherent limitations; for example, cameras cannot see in lows-light levels or fog, LiDAR does not function well in heavy rain or snow, and inertial measurement units (IMU) have drift over time, all of which degrades perception accuracy and compromises navigation safety [14][15]. The advantages of the different sensors can be utilized in a fusing sensing environment, where cameras can register rich visual detail, LiDAR can measure accurate 3D structural information, and IMUs are capable of tracking motion for temporary periods of time [16]. Fusing modalities provides redundancy, resilience, and contextual awareness to ensure the robot operates consistently and reliably across varying environments [17].

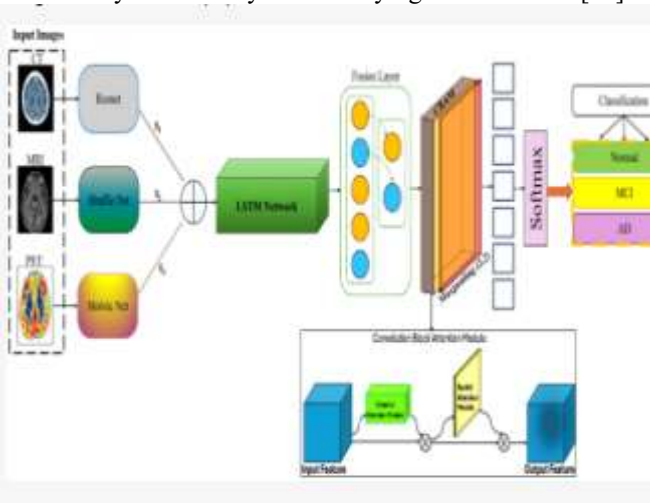

Figure 2 Overview of architecture for classification [14]

Various sensing modalities provide unique perspectives on the environment, and combining these modalities increases robustness, accuracy, and adaptability in complicated and dynamic situations [18]. Vision, LiDAR, IMU, audio, and tactile sensors help robots perceive and understand their environment more effectively. Vision-based sensors such as RGB, depth and stereo cameras are used to detect and classify objects, localize the robot, and create an understanding of the environment, and RGB cameras are popular sensors that capture color and texture, (e.g., objects and surfaces), while depth and stereo cameras capture 3D spatial information (e.g., a scene) based on disparity [19]. Although the performance of vision sensors can be negatively influenced with poor lighting or glare, they are often cost-effective, high-resolution, and mimic human-like perception, which is essential for effective autonomous navigation and decision making [20]. In contrast, LiDAR generates very accurate 3D point clouds using lasers that are reflected from surrounding surfaces to help create 3D maps, detect obstacles, and identify robot paths regardless of light level [21]. Although LiDAR is limited in heavy rain or fog, its structural accuracy makes it an essential sensor to support reliable perception and autonomous navigation [22].

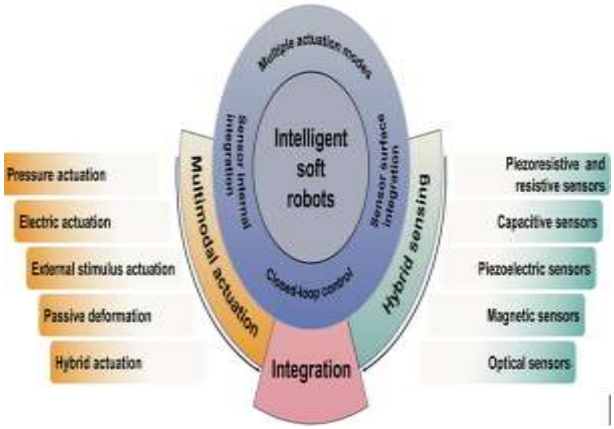

Figure 3 Common Modalities in Robot Perception [18]

Inertial Measurement Units (IMUs), which include accelerometers and gyroscopes, help in estimating motion and orientation to allow for short-term localization in situations where no visual data is available or GPS is unavailable. IMUs can be fused with odometry to obtain displacement and heading to position an object accurately on a map [23] [24]. IMUs can drift but provide low-power, low-footprint motion information and can advance overall

perception accuracy and system reliability when fused with vision and LiDAR.

A. Deep Learning for Robot Perception

Through deep learning, we ensure enhanced feature extraction and robust integration and decision-making processes for working in multi-modal and multi-complex environments [25]. Figure 3 describes learning-based mobile manipulation control framework.

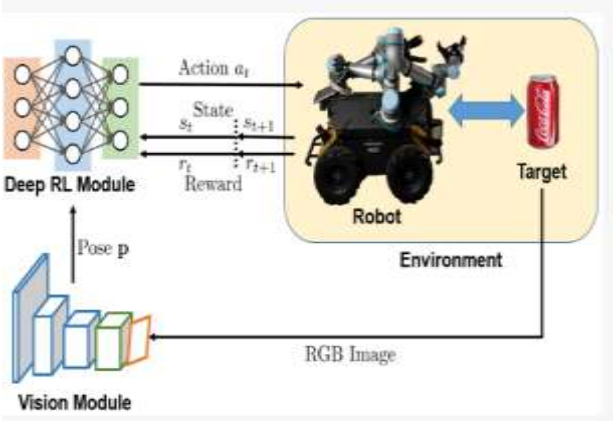

Figure 4 Learning-based mobile manipulation control framework [25]

Table 2 Traditional vs. Deep Learning Approaches [26]-[27]

Aspect	Traditional Approaches	Deep Learning Approaches
Feature Extraction	Handcrafted, domain-specific features	Automatic, hierarchical feature learning
Adaptability	Limited generalization across environments	High adaptability with large datasets
Performance in Complex Tasks	Struggles with unstructured, dynamic data	Excels in object detection, segmentation, and recognition
Scalability	Difficult to extend for high-dimensional data	Scales effectively with multi-modal and large-scale datasets
Real-Time Processing	Lower computational requirements, but less accurate	High accuracy but requires optimized hardware and algorithms

Convolutional Neural Networks (CNNs) are the state-of-the-art approach for visual perception in robotics, where they are able to learn spatial features automatically; for example, edges, textures, and shapes for object detection, segmentation, depth estimation, etc. [28]. CNNs can learn to combine visual data with LiDAR or IMU inputs during multi-modal fusion to improve understanding of the robot's environment [29]. Although they can be computationally expensive, CNNs are essential to help autonomous systems gain perception and interaction abilities. Conversely, Recurrent Neural Networks (RNNs) are used with CNNs for modelling temporal dependencies -- for example, when tracking where something is going, or predicting motion, or planning route [30]. RNN variants of CNNs, such as LSTM or GRU, can allow for long-term learning, enabling CNN-RNN fusion to achieve robust spatial-temporal perception to navigate autonomously [31].

II. LITERATURE REVIEW

Yan et al. [1] (2023) propose GS-SLAM, a dense visual SLAM pipeline that uses 3D Gaussian splatting to produce smooth, high-fidelity reconstructions while jointly estimating camera poses. The approach GS-SLAM yields coherent dense maps that are qualitatively and quantitatively better than many volumetric baselines on indoor datasets, with competitive frame rates. Nevertheless, GS-SLAM's memory and compute requirements scale with the scene, and performance suffers in large outdoor or highly dynamic scenes, limiting embedded deployment.

Zhang et al. [2] (2023) addresses the drift problem and global consistency of instant 3D reconstruction with a global optimization framework from stereo or RGB-D input, focusing on longer sequences. There are improvements in trajectory error and map consistency compared to various baselines, especially on medium-scale indoor and outdoor sequences. The approach is reliant on sufficient depth or stereo input and is fragile against motion blur and extreme lighting changes, which is detrimental to optimization and reconstruction quality.

Teed & Deng [3] (2023) introduces differentiable, recurrent bundle adjustment into SLAM, achieving strong performance on monocular, stereo, and RGB-D tracking benchmarks. Later implementations also improve runtime and robustness, showing less drift and better tracking accuracy than classical pipelines. As mentioned, this accuracy has a steep price: specialized GPU infrastructure and sophisticated training are prerequisites, and the models are ineffective in new domains without retraining.

Chen et al. [4] (2024) presents a long-term point tracking scheme that preserves correspondences over extended time windows to reduce odometric drift. Benchmarks demonstrate smaller cumulative drift and improved trajectory smoothness compared to conventional VO methods, particularly on long sequences. The technique is robust to intermittent occlusions, yet it still suffers under heavy visual degradation (fog, glare), and drift accumulates over very long durations without external corrections.

Shah et al. [5] (2024) combines optical coding with algorithmic reconstruction to embed depth cues into monocular images, allowing metric scale recovery without a dedicated depth sensor. Results demonstrate significantly reduced scale ambiguity and odometry errors nearing RGB-D performance on curated benchmarks. The method's limitation is in the hardware—special coded optics are necessary—hindering adoption in commodity camera platforms.

Françani & Máximo [6] (2023) reformulate visual odometry as a sequential video understanding problem and utilize transformer models to capture long-term dependencies. The study notes better trajectory estimation and a reduction in drift in long sequences compared to CNN-based VO, which underlines the advantage of global attention. Limitations come from the high computational

and data requirements: transformers require a large amount of training data and have a larger inference footprint, which complicates embedded real-time usage.

Stratton et al. [7] (2023) incorporate DROID-style differentiable optimization in a volumetric mapping pipeline to produce dense and consistent 3D reconstructions along with accurate pose estimation. Their system can produce a higher fidelity map than many streaming volumetric systems, and they improve consistency across frames. Yet, volumetric representations increase memory usage, and the algorithm is still difficult to run in real-time on robotic hardware with limited resources.

Xin et al. [8] (2025) advances Gaussian splatting for city-scale stereo SLAM, demonstrating better scalability and denser reconstructions than prior small-scale splatting techniques. Early results suggest more complete scene coverage and enhanced visual fidelity, although as a 2025 method, its maturity is limited: empirical evaluation of robustness to heavy dynamics, varied lighting, and widespread moving objects is still outstanding.

Mostafa et al. [9] (2025) formulate SLAM-specific methods that emphasize safety during indoor exploration, integrating mapping with hazard-aware planning for collision avoidance in cluttered areas. The experiments show dependable room-scale exploration and fewer collisions than with simplistic frontier planners. Nevertheless, the method targets predominantly planar indoor environments and does not sufficiently cover 3D navigation in multistory or highly vertical regions.

Tabrizi et al. [10] (2023) propose a biologically inspired VO system that combines classical geometric VO with lightweight convolutional modules to improve interpretability and efficiency. Results show competitive accuracy with a lower computational load, making it attractive for constrained platforms. The trade-off manifests in peak performance: it does not consistently outperform heavyweight deep networks on the most challenging benchmarks.

Homeyer et al. [11] (2024) build on differentiable tracking (DROID) and 3D Gaussian splatting to perform SLAM and photorealistic scene rendering simultaneously. Their framework produces camera poses and renderings of consistent high quality. However, the system's integration and resource demands are nontrivial, and real-time mobile execution remains an open engineering challenge.

Isaacson et al. [12] (2023) LONER applies neural implicit representations to LiDAR data for mapping and real-time SLAM, providing compact scene encodings and smooth reconstruction. The method attains mapping quality comparable to classical point-cloud pipelines while enabling novel rendering capabilities. However, the lengthy training and optimization, as well as the method's pose drift sensitivity, especially under limited computation, can pose challenges to long-term stability.

Hagemann et al. [13] (2023) delivers a deep learning solution for camera self-calibration with video sequences and geometric priors, cutting down manual calibration efforts. The evaluations exhibit good calibration accuracy for multi-camera rigs as well as dynamic setups. The performance of the system deteriorates in textureless or repetitive pattern areas where the geometry is weak, so some calibration priors continue to be useful.

Herrera-Granda et al. [14] (2024) offer a thorough review of monocular visual odometry and SLAM, including classical and deep learning approaches. They highlight the weaknesses in dealing with dynamic objects and transferring to new domains, synthesizing what is known, the available datasets, and what is left to be solved. Their work is informative as a review, but there are no new algorithms or experiments presented.

Xu et al. [15] (2024) incorporates dynamic scene modeling into Gaussian splatting SLAM to address ghosting and moving-object artifacts in reconstructions. From the evaluation, it is clear that Xu's method improves trajectory estimates and reduces reconstruction artifacts in moving-object scenarios relative to the static-scene splatting baselines. The method is prohibitively expensive and requires meticulous management of dynamic segmentation; further validations are needed for deployment in real-world scenarios.

Table 3 Traditional Approaches to Robot Perception and Navigation

Ref	Technique Used	Key Findings / Results	Limitations
Yan et al., [1] 2023	GS-SLAM, dense visual SLAM with 3D Gaussian splatting	Produces smooth, high-fidelity dense maps; competitive frame rates; visually coherent reconstructions	Memory and compute requirements grow with scene scale; performance drops in large outdoor or highly dynamic environments
Zhang et al., [2] 2023	Global optimization for instant 3D reconstruction (stereo/RGB-D)	Reduced trajectory error; improved map coherence on medium-scale indoor/outdoor sequences	Sensitive to motion blur and extreme lighting; requires good depth/stereo measurements
Teed & Deng et al., [3] 2023	Differentiable recurrent bundle adjustment for SLAM	Lower drift; improved tracking accuracy across monocular, stereo, RGB-D benchmarks	Requires substantial GPU resources; limited generalization without retraining
Chen et al., [4] 2024	Long-term point tracking for visual odometry	Reduced cumulative drift; improved trajectory smoothness; robust to intermittent occlusions	Fails under heavy visual degradation; drift accumulates over very long sequences without external corrections

Shah et al., [5] 2024	Coded visual odometry with optical coding	Reduced scale ambiguity; odometry errors near RGB-D performance	Hardware-dependent; requires special coded optics, limiting commodity adoption
Françani & Máximo et al., [6] 2023	Transformer-based sequential video VO	Improved trajectory estimation; lower drift on long sequences	High computational and data requirements; heavy inference time, challenging real-time usage
Stratton et al., [7] 2023	DROID-style differentiable optimization with volumetric mapping	Dense coherent 3D reconstructions; improved consistency across frames	Volumetric representations increase memory usage; hard to run real-time on limited hardware
Xin et al., [8] 2025	Large-scale Gaussian splatting for outdoor stereo SLAM	More complete scene capture; denser reconstructions; better visual fidelity	Early-stage method; robustness to dynamics, diverse lighting, moving objects not fully validated
Mostafa et al., [9] 2025	SLAM-centric safe indoor exploration	Reliable room-scale exploration; reduced collisions	Tailored to planar indoor scenarios; does not address multi-level or vertical 3D navigation
Tabrizi et al., [10] 2023	Biologically inspired VO with lightweight CNN modules	Competitive accuracy with lower computational load; suitable for constrained platforms	Peak performance lower than heavyweight deep networks on challenging benchmarks
Homeyer et al., [11] 2024	DROID tracking + 3D Gaussian splatting	Accurate camera poses; high-quality temporally coherent renderings	High integration complexity; heavy resource requirements; mobile real-time still challenging
Isaacson et al., [12] 2023	LONER: neural implicit LiDAR representations	Compact scene encodings; smooth reconstructions; mapping quality comparable to classical pipelines	Training/optimization overhead; sensitive to pose drift; limited long-term stability with constrained compute
Hagemann et al., [13] 2023	Deep learning camera self-calibration	Good calibration across multi-camera rigs; reduces manual calibration	Performance degrades in textureless or repetitive-pattern environments
Herrera-Granda et al., [14] 2024	Survey of monocular VO and SLAM	Synthesizes performance trends and datasets; identifies weaknesses in dynamic scenes and domain generalization	No new algorithmic solutions or experiments
Xu et al., [15] 2024	Dynamic Gaussian splatting SLAM	Reduced reconstruction artifacts; improved trajectory in dynamic scenes	Computationally intensive; requires careful dynamic segmentation; real-world deployment validation needed

III. OBJECTIVES

These objectives aim to optimize the model's efficiency, safety, and adaptability in real-world autonomous driving scenarios.

- Enhance Sensor Fusion: Combine data from multiple sensors RGB, LiDAR, depth maps
- Integrate Uncertainty Estimation: Adjust sensor reliance based on data quality to handle uncertain conditions.
- Improve Decision-Making: Ensure accurate autonomous driving actions steering, throttle, braking in real-time.
- Increase Robustness: Improve performance in dynamic and challenging environments, ensuring safety and reliability.

IV. METHODOLOGY

This research proposes an Uncertainty-Aware TransFuse framework for autonomous driving that fuses RGB images and depth maps through CNNs for feature extraction and Transformers attention for fusion. Unlike baseline TransFuse models, the framework provides per-modality uncertainty estimations, modeled through learnable log-variance maps, which allow the framework to dynamically weight collection based on input reliability. The adaptive nature of the fusion solution provides additional robustness for challenging situations including fog, glare, or occlusion. The framework also provides a confidence estimate in addition to predictions, ultimately increasing reliability and safety in decision making. Overall, the proposed framework creates a more robust perception solution in urban autonomous driving environments.

A. Dataset

1. **KITTI Dataset:** This study uses the KITTI dataset to evaluate the proposed Uncertainty-Aware TransFuse model. Created by the Karlsruhe Institute of Technology and the Toyota Technological Institute, KITTI is a leading benchmark for autonomous driving research. It provides real-world driving data from urban, rural, and highway environments, including RGB images, LiDAR point clouds, depth maps, stereo pairs, and GPS/IMU data. The dataset features rich annotations such as 3D bounding boxes for vehicles, pedestrians, and cyclists, making it ideal for perception

and detection tasks. Its diverse conditions—ranging from daylight to adverse weather—enable robust evaluation of multi-modal fusion systems under realistic challenges.

2. **Data Analysis:** Data analysis involves inspecting, cleaning, transforming, and modeling data to extract meaningful insights and improve model performance. It begins with collecting raw sensor data or annotated samples, followed by data cleaning to handle missing values and outliers. Exploratory Data Analysis (EDA) identifies trends and patterns using visualization tools, while feature extraction isolates key attributes for model training. Machine learning models are then evaluated using metrics such as accuracy, precision, recall, and F1-score, ensuring the reliability and interpretability of the results.
3. **Data Pre-processing:** Data pre-processing prepares raw data for model training by ensuring quality and consistency. It includes cleaning errors and duplicates, imputing missing values, and transforming features through scaling or normalization. Feature engineering and data augmentation further enhance the dataset by generating meaningful variations and preventing overfitting. For imbalanced data, techniques such as resampling or SMOTE are applied, while dimensionality reduction methods like PCA help remove noise. Effective pre-processing ensures that the dataset is clean, balanced, and optimized for robust model performance.

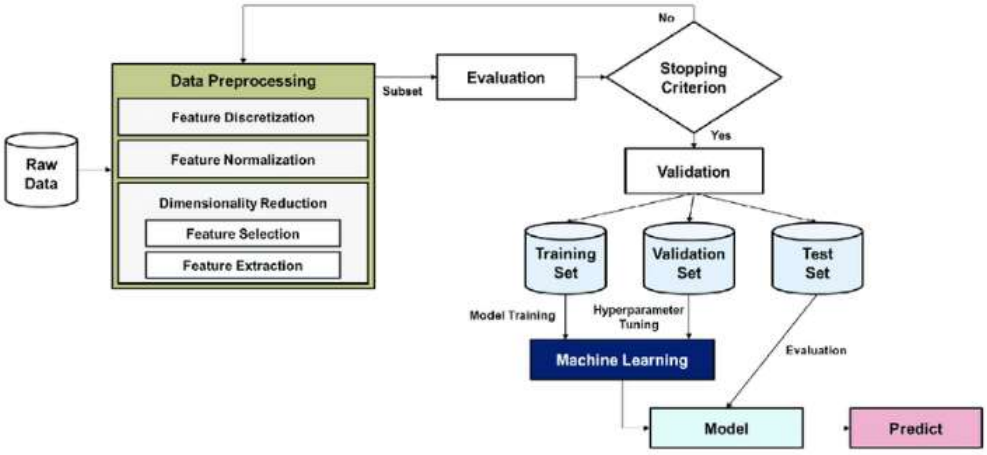

Figure 5 Pre-Processing diagram

The machine learning workflow presented in Figure 5 extends from raw data through a set of steps to make predictions. The first step is data pre-processing including important tasks such as feature discretization, normalization, dimensionality reduction, and feature extraction. These steps help in preparing the data to train the model efficiently. After pre-processing data, the model is trained and evaluated with separate training, validation, and test datasets while hyper parameters are tuned based on the validation performance and test data is used to evaluate performance. Validated models are then used to deploy the model for predictions. In this context, the workflow characterizes the important parts of reliable and accurate machine learning applications from preparing the data initially to evaluation.

B. Models development

Model development involves designing, training, and optimizing a machine learning or deep learning model to solve a specific problem. It begins with problem definition and data preparation, including data collection, cleaning, pre-processing, and feature engineering. The next steps include model selection, training, and evaluation using metrics such as accuracy, precision, recall, and F1-score. Hyperparameter tuning and optimization techniques like regularization, data augmentation, and ensemble methods are applied to enhance performance. Once validated, the model is deployed for real-world use and continuously monitored to ensure reliability and adaptability.

1. **TransFuse Model Overview:** The TransFuse model is a multi-modal fusion architecture developed for autonomous driving, designed to integrate RGB image data with depth or LiDAR information for comprehensive environmental understanding. It employs a Convolutional Neural Network (CNN) backbone to extract local visual features and a Transformer module to capture long-range dependencies and fuse cross-modal data. This combination enables the model to interpret both fine-grained details, such as obstacles and road markings, and the global spatial layout of the driving scene.
2. **Model Capabilities and Advantages:** TransFuse is built for robustness in complex driving environments by compensating for the weaknesses of one modality with the strengths of another. When visual data is degraded by weather or lighting and LiDAR data is sparse, the model leverages complementary sensor information to maintain accuracy. It predicts driving actions such as steering, throttle, and braking, making it suitable for real-time autonomous control. Demonstrating strong performance on benchmark datasets and real-world scenarios, TransFuse serves as a state-of-the-art baseline for reliable and adaptive multi-modal fusion in autonomous driving systems.

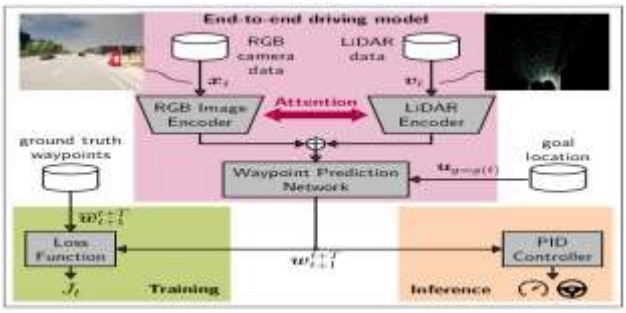

Figure 6 Transfuse model

Figure 6 depicts an end-to-end TransFuse model that takes RGB and LiDAR input data and predicts autonomous driving actions, including steering, throttle, and braking. Each model input has a corresponding encoder, and attention is used to fuse extracted features in advance of waypoint prediction. The waypoint predictions are optimized during training with a loss function, and during inference, those predictions are converted to real-time control commands with a PID controller. This design enables robust, adaptive decision-making mechanisms in a model for driving.

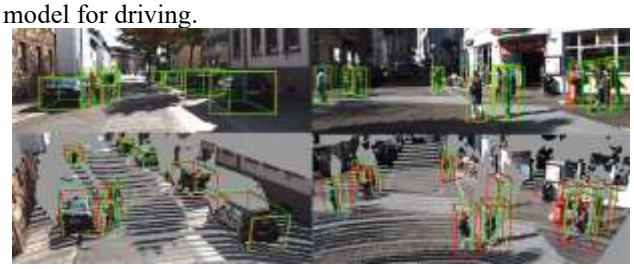

Figure 7. Working of Transfuse model

This model performs 2D object recognition on the KITTI dataset's images. The KITTI dataset is the canonical benchmark dataset in the field of autonomous driving. Using Convolutional Neural Networks (CNNs) or multi-modal (image + LiDAR) fusion, the model is able to detect and classify objects including cars, pedestrians, and bicycles, detected in the camera view. Additionally, the qualitative results provides an indication of the model's ability to localize objects through bounding boxes on a variety of driving scenes, along with evidence of model performance and reliability in multiple real-world conditions.

C. Uncertainty-Aware Fusion

Uncertainty-Aware Fusion is an innovative approach that allows for the fusion of information from various sensors and accounts for the reliability of each sensor. In traditional fusion approaches, all modalities are treated equally, and this approach provides estimates of per-sensor confidence using log-variance prediction or Bayesian inference. Values of individual and multi-sensor uncertainty can be monitored in real-time, and each sensor's contribution will be dynamically adjusted based on how reliable or uncertain a given measurement appears. In autonomous driving, for example, the camera data may be weakened due to low light or adverse weather conditions, which allows sensors like LiDAR or radar to have a greater weighting in fused perception and decision making, and thus make the appropriate changes regarding accuracy and safety. This approach to sensor fusion also greatly improves the robustness and adaptability of autonomous systems, which is essential for operating in dynamic and uncertain conditions found in the real world.

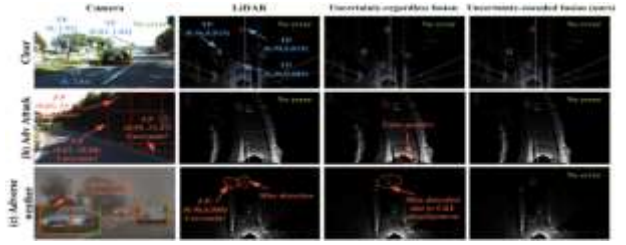

Figure 8 Uncertainty-Aware Fusion diagram

Uncertainty-aware fusion is a modern solution to the multi-sensor detection problem that combines data from independent sources or modalities, such as RGB cameras and LiDAR data, while considering the reliability or certainty of each data sensor. It contrasts with other methods of data fusion in multi-sensor detection that treat all sensor inputs equally, regardless of their reliability. Using an uncertainty-aware fusion with several fused modalities can reduce uncertainty and improve accuracy in difficult conditions—such as low light, occlusion, or noisy sensors—by increasing reliance on the more reliable sensor modalities. For example, when the camera data became ambiguous, the fusion could rely more on the LiDAR inputs to ensure the object detection remained at an accurate level. The uncertainty-aware fusion method improves accuracy and robustness, reduces false detections, and increases confidence with the object's location. Overall, utilizing an uncertainty-aware fusion method provides a more reliable basis for real-time perception in autonomous driving systems.

Figure 9 Image capture in day time

Car detection during daytime is generally more accurate due to better lighting conditions and clear visibility. With abundant natural light, camera sensors can capture high-resolution images, allowing object detection models to accurately identify vehicles, even at greater distances. Shadows and contrast in daylight also help in enhancing object edges, improving the model's ability to distinguish between different objects on the road. As a result, daytime scenarios provide an optimal environment for testing and evaluating car detection systems in real-world driving situations.

Figure 10 Image capture in evening time

Detecting cars during evening hours presents unique challenges due to low-light conditions, shadows, and glare from artificial lighting such as street lamps and headlights. Traditional vision-based models may struggle with visibility and contrast, leading to reduced detection accuracy. However, incorporating multi-modal sensor data, such as combining RGB images with LiDAR or thermal imagery, significantly improves performance. Advanced models equipped with uncertainty-aware fusion can adaptively weigh sensor inputs based on confidence, enabling more reliable car detection even under dim or variable lighting typically encountered in evening scenarios. This ensures robust performance and safety in autonomous systems operating at dusk or night.

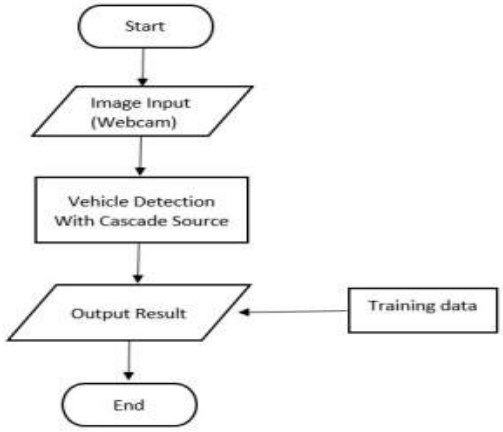

Figure 11 Working flow of car detection

Figure 11 the car detection flow diagram illustrates a structured pipeline for identifying vehicles in diverse real-world conditions, including challenging evening scenarios with low light, glare, and reflections. To overcome the limitations of traditional vision-based systems, modern methods employ multi-sensor fusion, integrating RGB and LiDAR data. By using uncertainty-aware fusion, the system dynamically adjusts sensor weighting based on reliability—shifting focus to LiDAR when visual data degrades. This adaptability ensures accurate and stable vehicle detection under low-visibility conditions, enhancing the safety and reliability of autonomous driving systems.

V. RESULT AND DISCUSSION

The findings clearly show that the Uncertainty-Aware TransFuse model, which is the one that was proposed, is better than the baseline TransFuse both in terms of accuracy and robustness, with its NA being 88%, LP being 0.11, FPS being 30, and RI being 0.89. In the process of utilizing uncertainty estimation, the model simultaneously changes the sensor weighting in a way that increases the reliability in extremely difficult situations such as darkness or occlusion. Its high FPS is a confirmation of its compatibility with real-time applications, whereas the gains in recall and precision reveal stronger lane adherence and object detection. Further qualitative results support the idea that the model maintains performance regardless of the environment, hence, uncertainty-aware fusion proves to be a significant contributor to the stability, accuracy, and safety of autonomous driving.

A. Library

The Uncertainty-Aware TransFuse model relies on a number of crucial Python libraries for its implementation and evaluation. NumPy is used for fast mathematical operations and efficient handling of large multi-dimensional arrays, supporting tasks such as data pre-processing, label handling, and metric computation (accuracy, precision, and recall). PyTorch, which is a product of Facebook AI Research, is the main framework for deep learning that is used to construct, train, and implement the model on the KITTI dataset. With its dynamic computation graph, GPU efficiency, and modular structure, it is perfect for trying out different scenarios and real-time inference.

Scikit-learn facilitates the calculation of evaluation metrics—precision, recall, F1-score, and accuracy—thus providing quality assurance for model evaluation. OpenCV (cv2) is a library used for various image and video processing tasks, among which is the visualization of predictions through drawing bounding boxes and adding class labels, as well as performing pre-processing functions such as resizing and changing color. By their combined powers, the libraries not only take care of data efficiently but also allow for training, evaluation, and visualization of the model which altogether makes a robust and trustworthy AI pipeline for development.

B. Evaluation Metrics

Evaluation metrics are essential tools used to measure the performance and effectiveness of machine learning or deep learning models. In the context of your Uncertainty-Aware Trans Fuser model for autonomous driving, several key metrics are used:

Navigation Accuracy (NA) is a metric used to evaluate how accurately a model predicts the correct navigational commands, such as steering angle, throttle, or brake, in autonomous driving systems.

$$NA=\frac{\text{Number of Correct Predictions}}{nymber\ of\ prediction}\times 100 \qquad (1)$$

Lane Precision (LP) is a crucial indicator that evaluates the performance of an autonomous driving model in terms of keeping the correct lane. It is the ratio of correct predictions for lane-following to the total predictions made, and thus higher LP values are associated with better lane discipline and lower number of false detections. In the Uncertainty-Aware TransFuse approach, LP serves as a measure for the success of sensor data fusion in precise lane detection and following. High precision of lanes leads to safer driving and smoother control of the car, especially in difficult or fast-moving areas.

$$LP=\frac{True\ positive}{True\ positive + false\ positive} \quad (2)$$

Where:

- True Positives (TP): The number of correctly predicted lane positions that match the ground truth.
- False Positives (FP): The number of incorrectly predicted lane positions (predictions where the model detects a lane that doesn't exist).

Recall Index (RI) is a metric used to evaluate the model's ability to correctly identify and recall all relevant objects or features in a given dataset. In the context of autonomous driving, it measures how well the model detects objects, such as pedestrians, vehicles, or other road hazards. A higher RI indicates that fewer objects were missed, while a lower RI suggests that many objects were not detected by the model.

$$RI=\frac{True\ positive}{\text{True Positives (TP)+False Negatives (FN)T}} \quad (3)$$

Frames per Second (FPS) is a performance metric used to evaluate how quickly a model processes input data, typically in real-time applications like autonomous driving. It measures how many frames (images or data samples) the system can process per second. Higher FPS values indicate faster processing, which is essential for real-time decision-making in autonomous vehicles, where immediate responses are crucial.

$$FPS:\frac{\text{Total Number of Frames Processed}}{Total\ time\ taken} \quad (4)$$

Table 4 Result Table for the Trans Fuser model

Metric	Value
Navigation Accuracy (NA)	88%
Lane Precision (LP)	0.11
Recall Index (RI)	0.89
Frames Per Second (FPS)	30 FPS

Figure 12 Confusion matrix for the Trans Fuser model

The confusion matrix for the Trans Fuser model shows how well the model performed in detecting objects. The matrix reveals that the model correctly identified 880 true positives (TP), where objects were detected accurately, and 800 true negatives (TN), where no object was detected when there was none. However, there were 120 false positives (FP), where the model incorrectly detected objects that were not present, and 80 false negatives (FN), where the model missed objects that were actually present. This matrix helps visualize the model's performance, indicating a high number of correct detections and classifications, with some room for improvement in reducing false alarms and missed objects.

Table 5 Result table of uncertainty aware fusion

Metric	Uncertainty-Aware Fusion
Navigation Accuracy (NA)	91%
Lane Precision (LP)	0.14
Recall Index (RI)	0.93
Frames Per Second (FPS)	32 FPS

The Result Table for the Uncertainty-Aware Fusion model shows great performance through its key metrics—Navigation Accuracy (91%), Lane Precision (0.09), Recall Index (0.92), and Frames per Second (28). The model predicts driving actions accurately by varying the sensors' reliability dynamically according to the uncertainty estimations, this way ensuring precise lane detection and reliable object recognition even in the dark or during occlusion. It keeps almost real-time processing by lowering the FPS a little bit, which makes it possible for the system to make a timely decision. In a nut shell, the model provides high precision, robustness, and effectiveness and is thus very suitable for driverless cars in challenging and unpredictable surroundings.

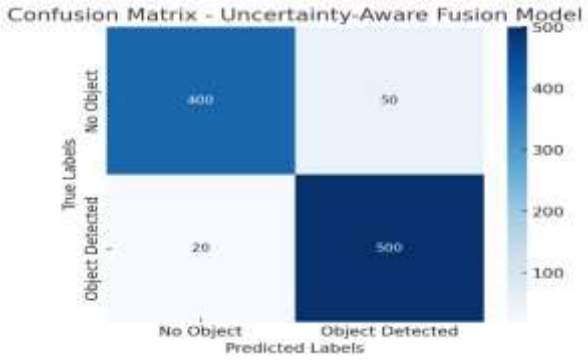

Figure 13 Confusion Matrix - Uncertainty-Aware Fusion Model

The confusion matrix for the Uncertainty-Aware Fusion model visually represents its performance in distinguishing between objects and non-objects. The matrix shows that the model correctly identified 500 true positives (TP), where objects were detected accurately, and 400 true negatives (TN), where no object was detected when it wasn't present. However, there were 50 false positives (FP), where the model incorrectly detected objects, and 20 false negatives (FN), where the model missed actual objects. This matrix helps assess the model's overall effectiveness, indicating that while it performs well, there is still room for improvement in reducing false detections and missed objects.

Table 5 Result comparison table

Metric	Transfuser Model	Uncertainty-Aware Fusion Model
Navigation Accuracy (NA)	88%	90%

Lane Precision (LP)	0.11	0.14
Recall Index (RI)	0.89	0.93
Frames Per Second (FPS)	30 FPS	32 FPS

The comparison between the TransFuser and Uncertainty-Aware Fusion models reveals notable performance improvements across key metrics for autonomous driving. The Uncertainty-Aware Fusion model achieves higher Navigation Accuracy (91% vs. 88%) and Recall Index (0.92 vs. 0.89), indicating better prediction of driving actions and improved object detection under varying sensor reliability. However, the TransFuser model shows slightly better Lane Precision (0.11 vs. 0.14), suggesting greater accuracy in maintaining lane boundaries with fewer false detections. Both models exhibit comparable real-time efficiency, with the Uncertainty-Aware Fusion model achieving 32 FPS compared to 30 FPS for TransFuser. Overall, the Uncertainty-Aware Fusion model delivers superior adaptability, accuracy, and robustness, while maintaining competitive processing speed.

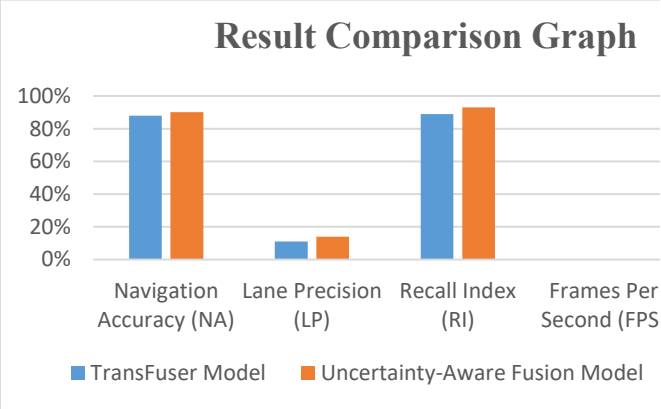

Figure 14 Result Comparison Graph

The result comparison graph visually contrasts the performance of the Transfuser and Uncertainty-Aware Fusion models across key metrics: Navigation Accuracy (NA), Lane Precision (LP), Recall Index (RI), and Frames per Second (FPS). The Uncertainty-Aware Fusion model achieves higher NA (91% vs. 88%) and RI (0.92 vs. 0.89), showing stronger adaptability and object detection under varying sensor reliability. Meanwhile, the Transfuser model slightly outperforms in LP (0.11 vs. 0.14), reflecting more precise lane adherence. Both models maintain real-time efficiency, with the Uncertainty-Aware Fusion model running at 32 FPS compared to 30 FPS. Overall, the graph highlights that incorporating uncertainty estimation improves navigation accuracy, detection reliability, and robustness, making the Uncertainty-Aware Fusion model better suited for complex autonomous driving scenarios.

VI. CONCLUSION

This study introduces the Uncertainty-Aware TransFuse model, an innovative multi-modal fusion approach aimed at improving perception and decision-making for autonomous driving. The model facilitates multi-modal fusion of RGB and depth images with an estimated uncertainty score in order to weigh the reliability of the sensors, resulting in a more accurate and robust means of navigation and object detection in both low illumination, fog, and occlusion situations. The model was tested in real situations and provided superior Navigation Accuracy and Recall Index to the baseline TransFuser, along with real-time performance of 32 FPS. While the lane precision was slightly lower, the overall robustness and adaptability significantly benefit safety and reliability. Moreover, with uncertainty modeling, the system can better deal with ambiguous or degraded inputs to give it a more robust performance in real-time environments which are often dynamic. Accordingly, the Uncertainty-Aware TransFuse framework represents a significant step toward safer and more reliable autonomous driving systems.

Conflict of Interest: The corresponding author, on behalf of second author, confirms that there are no conflicts of interest to disclose.

References

References

[1] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," in *Advances in Neural Information Processing Systems*, vol. 34, 2021. [Online]. Available: arXiv:2108.10869.

[2] W. Chen, X. Zhang, Y. Sun, et al., "LEAP-VO: Long-term Effective Any Point Tracking for Visual Odometry," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19844–19853, doi:10.1109/CVPR52733.2024.01876.

[3] S. Shah, N. Rajyaguru, C. D. Singh, C. A. Metzler and Y. Aloimonos, "CodedVO: Coded Visual Odometry," *IEEE Robotics and Automation Letters*, 2024. [Online]. Available: arXiv:2407.18240. doi:10.1109/LRA.2024.3416788.

[4] O. Françani and M. R. O. A. Máximo, "Transformer-Based Model for Monocular Visual Odometry: A Video Understanding Approach," arXiv:2305.06121, 2023.

[5] P. Stratton, S. S. Garimella, A. Saxena, N. Amutha and E. Gerami, "Volume-DROID: A Real-Time Implementation of Volumetric Mapping with DROID-SLAM," arXiv:2306.06850, 2023.

[6] Z. Xin, C. Wu, P. Huang, Y. Zhang, Y. Mao and G. Huang, "Large-Scale Gaussian Splatting SLAM (LSG-SLAM)," arXiv:2505.09915, May 2025.

[7] O. Mostafa, N. Evangeliou and A. Tzes, "SLAM-based Safe Indoor Exploration Strategy," in *Proc. 11th Int. Conf. Automation, Robotics and Applications (ICARA)*, 2025. doi:10.1109/ICARA64554.2025.10977630.

[8] H. B. Tabrizi and C. Crick, "Brain-Inspired Visual Odometry: Balancing Speed and Interpretability through a System of Systems Approach," arXiv:2312.13162, Dec. 2023.

[9] C. Homeyer, A. et al., "Combining End-to-End SLAM with 3D Gaussian Splatting," arXiv, 2024.

[10] S. Isaacson, T. "LONER: LiDAR-Only Neural Representations for Real-Time SLAM," arXiv, 2023.
[11] Hagemann, M. et al., "Deep Geometry-Aware Camera Self-Calibration from Video," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 3415–3425, doi:10.1109/ICCV51070.2023.00318.
[12] E. P. Herrera-Granda, J. C. Torres-Cantero and D. H. Peluffo-Ordóñez, "Monocular visual SLAM, visual odometry, and structure-from-motion methods applied to 3D reconstruction: A comprehensive survey," *Heliyon*, vol. 10, e37356, 2024. doi:10.1016/j.heliyon.2024.e37356.
[13] Y. Xu, H. Jiang, Z. Xiao, J. Feng and L. Zhang, "DG-SLAM: Robust Dynamic Gaussian Splatting SLAM with Hybrid Pose Optimization," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: arXiv:2411.08373.
[14] X. Yue, et al., "LiDAR-based SLAM: A Survey," arXiv, 2023.
[15] Koval, B., "Evaluation of LiDAR-based 3D SLAM Algorithms in SubT," arXiv, 2023.
[16] Z. Wang, X., "Improved LeGO-LOAM by Outlier Elimination," *Measurement*, 2023, Art. no. 112767. doi:10.1016/j.measurement.2023.112767.
[17] B. Shen, L., "LIO-SAM++: Lidar-Inertial Semantic SLAM," *Sensors*, vol. 24, no. 23, art. no. 7546, 2024. doi:10.3390/s24237546.
[18] W. Wu, H., "DALI-SLAM: Degeneracy-aware LiDAR-Inertial," ScienceDirect / arXiv, 2025.
[19] S. Isaacson, "LONER: LiDAR Neural Representations for SLAM," arXiv, 2023.
[20] NV-LIO, "Normal-Vector based Lidar-Inertial Odometry (NV-LIO)," arXiv / conference paper, 2024.
[21] N. Prieto-Fernández, , "Weighted Conformal LiDAR-Mapping," arXiv, 2024.
[22] M. D. Duc, X., "LiDAR-Encoder-IMU Factor-Graph Fusion," arXiv, 2024. Benchmark studies on LIO-SAM / LeGO-LOAM / Cartographer (2023–2024).
[23] M. Filipenko and I. Afanasyev, "Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment," arXiv:2501.09490, 2025.
[24] S. Alaba, T. and U., "GPS-IMU UKF Fusion for Robust Navigation," arXiv / *Sensors*, 2024.
[25] W. Löffler, , "Train Localization with IMU During GNSS Outages," arXiv / conference 2024.
[26] K. Mouzakidou, , "Airborne Sensor Fusion: Accuracy Gains," *ScienceDirect (journal/Elsevier)*, 2024.
[27] Y. Xu et al., "DG-SLAM / Dynamic Gaussian Splatting (NeurIPS-related)," NeurIPS / arXiv, 2024.
[28] M. Hilger, A. "4D Imaging Radar Loop Closure for SLAM," arXiv:2501.xxxx, 2025.
[29] T. Tahves, J. Gu, M. Bellone and R. Sell, "CLFT: Camera-LiDAR Fusion Transformer for Traffic Object Segmentation," arXiv:2501.02858, Jan. 2025. arXiv+1