Journal of Applied Automation Technologies

Research Paper

A Novel Deep Learning Approach for Multi-Sensor Fusion in Autonomous Vehicle Perception

Paulina Patrycja Królowa[1,*], Patryk Pacholski[1] and Oliwier Orzechowski[1]

[1] Faculty of Automatic Control, Robotics and Electrical Engineering, Poznan University of Technology, Poznan, 60-965, Poland
*Corresponding author: paulina.pk@put.poznan.pl

Abstract. An effective autonomous driving perception system must function well in real-world settings. In order to enhance situational robustness and environmental perception, LiDAR, radar, and camera data must be combined. This study examines the ongoing challenges in multi-sensor fusion. First, create a deep learning-based fusion framework that can manage the diverse spaces, timings, and semantics of the multiple sensors in a systematic manner. In order to implement attention-based fusion, adaptively extract features, and dynamically estimate uncertainty in the perception pipeline for context-aware decision-making, a new structure has been developed. Perform multi-stage attention weighting and cross-modal integration after methodically encoding and aligning each sensor stream separately. Experiments using a large public dataset have demonstrated that the suggested approach is more suited for real-world autonomous driving scenarios. The new framework is still a real-time system with minimal latency and an inference speed of 27 frames per second; quantitatively, it has improved the mean Average Precision (mAP) by more than 6 percentage points. To make sure that multiple item tracking and detection remain accurate, robustness has been evaluated in inclement weather and sensor deterioration. In summary, this study has offered a comprehensive and workable solution to the perception issue in intelligent vehicles, and experimental results have demonstrated its effectiveness, flexibility, and potential use in urban traffic scenarios.

Keywords: *Sensor Fusion, Autonomous Driving, Deep Learning, Robust Perception*

Received on 29 December 2024, Accepted on 08 June 2025, Published on 16 June 2025

Introduction

A In recent years, autonomous driving has advanced quickly to increase traffic efficiency and road safety while also giving individuals new mobility options. A sophisticated vision system that can reliably and precisely identify both fixed and moving objects in a high-complexity, multi-environment situation is essential to the realisation of self-driving cars. In order to gather various kinds of environmental data for autonomous cars, LiDAR, radar, and high-resolution camera technology are now being combined in diverse ways [1]. In order to improve coverage, overcome the shortcomings of single sensors, and achieve full-featured situational awareness in challenging operating conditions, multi-sensor fusion has been introduced to create a strong-core perception system [2].

Nevertheless, there is yet no reliable and highly effective multi-sensor fusion-based perception system. The first is the heterogeneity of sensor data, which makes it challenging to extract and align mutually meaningful characteristics because of differences in data structure, geographic and temporal resolution, and susceptibility to environmental variables [3]. In fact, the system's performance can be readily decreased by asynchronous data streams, sensor noise, miscalibration, or partial sensor failure [4]. Fusion systems may produce duplicated, missing, or erroneous data since they should be steady in the face of inclement weather, occlusion, dynamic objects, and a changing environment [5]. Algorithmic complexity and model delay are also strictly regulated

because these are safety-critical applications that need to operate in real-time and have limited onboard computer resources [6]. While deep learning has improved multi-modal fusion's feature extraction efficiency [7], there are still challenges with generalisation, interpretability, and robustness, among other issues [8]. Universities and businesses have not yet addressed these issues [9].

In order to achieve high-performance perception in autonomous driving, this research proposes a novel deep learning architecture for multi-sensor fusion. Adaptive feature extraction has been offered as a novel approach to address the space-semantic mismatch of all-weather data. To increase the model's capacity for generalisation, attention-based fusion processes can contextually and flexibly integrate data from many sensors in various settings. The entire framework has been designed for high-speed inference and can operate on embedded systems due to its low memory and processing needs. The suggested approach has obtained the new best results in both detection accuracy and system efficiency based on several experiments carried out in the international benchmark datasets. This paper's primary contributions are: (1) an adaptive feature extraction architecture for heterogeneous sensor fusion; (2) a novel attention-based deep fusion module capable of context-aware integration; (3) a system design optimised for real-time embedded inference; and (4) comprehensive experimental verification results that establish new benchmarks for multi-sensor autonomous driving perception.

Related Work

Multi-Sensor Fusion Techniques in Autonomous Driving

The fundamental technique for perception in autonomous driving systems is multi-sensor fusion, which combines many types of sensors input to get a comprehensive picture of the surroundings [10]. Each of LiDAR, radar, and cameras has advantages and disadvantages of its own. For instance, LiDAR is highly accurate in 3D geometric structure, radar can detect velocity directly in inclement weather, and pictures offer rich semantic and textural information [11]. In general, early fusion, mid-level fusion, and late fusion are the three fundamental frameworks for integrating the aforementioned modalities [12].

In early fusion procedures, the network learns cross-modal characteristics from a source of raw or minimally processed sensor input. Significant issues in handling various spatial resolutions, sensor synchronisation, and signal-to-noise ratios arise from this method's excessive information content [13]. In order to balance flexibility and performance, mid-level fusion collects features individually from each sensor before performing concatenation or correlation in a shared latent space. However, this may result in the loss of fine-grained cross-modal interactions [14]. Conversely, late fusion employs a resilient and flexible system design and integrates high-level judgements or observed objects from modality-specific subnetworks, although it may lose some information in the intermediary levels [15]. The practical issues of precise time alignment, recalibration, sensor failure, and data loss in dynamic and crowded road settings affect all of the aforementioned techniques [16].

Although probabilistic sensor redundancy models and good alignment algorithms have been established in theory to address the aforementioned issues, scaling up these solutions continues to be a persistent engineering and research difficulty. Fusion design is a crucial and continuously developing field because industry deployment must also take into account real-time constraints, sensor placement, calibration drift, long-term maintenance, etc. [17].

Deep Learning Approaches for Sensor Fusion

Deep learning has revolutionised the model for sensor fusion in autonomous driving, leading to the emergence of end-to-end, data-driven solutions that outperform conventional rule-based techniques [18]. Voxel-based networks and set-based learning have also achieved direct operation on irregular 3D data, and convolutional neural networks are currently widely employed to extract spatial information from camera images and LiDAR point clouds [19]. Long-range dependencies have recently been modelled using transformer architecture, and multi-modal fusion can now use cross-modal attention to weight distinct sensor inputs differently in different driving contexts [20]. The flexibility and context sensitivity of fusion have been improved by attention-based and gating mechanisms to more effectively integrate redundant or complementary information in an adaptive manner [21].

The present models nevertheless need to be adequately built to prevent overfitting and guarantee steady operation in practice, notwithstanding significant developments in recent years. Researchers have steadily become interested in the effects of data augmentation, multi-task learning, and self-supervised approaches; these are currently widely used to address the issue of limited complete-annotated multi-modal datasets in the automotive area. Other outstanding technical issues include sample efficiency, scalability for edge devices, and model interpretability [22].

Due to the increased demand for deployment in commercial autonomous vehicles, the network must meet strict real-time and power requirements and have excellent detection and tracking accuracy. As a result, a lot of research has been done on robust yet lightweight deep-fusion models, uncertainty estimation approaches, and enhanced pipeline optimisation at the hardware and software levels.

Limitations of State-of-the-Art

Even if there have been some recent advancements, there are still several shortcomings in the deep learning frameworks and multi-sensor fusion techniques used today. Robustness in the face of extreme weather, rare obstacle configurations and sensor failures has not been achieved; most of the leading models are still limited to controlled benchmark datasets and perform poorly in open-world scenarios [23]. Generalization to unseen domains or different sensor setups is often limited by overfitting, lack of regulation, or insufficient calibration mechanisms [24].

Given the real-time constraints of embedded systems in automotive applications, it is difficult to achieve high-performance, low-latency deep fusion networks that are both memory-efficient and fast-inferencing [25]. Transparency and predictability of the model also need to meet regulatory and certification requirements for further study. Balance these four competing demands for accuracy, efficiency, interpretability and reliability, and further innovations in multi-sensor fusion technology for autonomous driving will continue to be needed.

Methodology

Adaptive Feature Extraction for Heterogeneous Sensors

Extracting informative and robust features from heterogeneous sensors is the first critical stage for high-precision autonomous driving perception. LiDAR, camera, and radar each present unique data characteristics: LiDAR generates sparse but highly accurate 3D point clouds; cameras provide dense two-dimensional images with rich semantic context; radar supplies versatile range-doppler maps, contributing notable resilience to adverse weather and lighting conditions. These variations create both opportunities for redundancy and complementariness, as well as substantial computational and calibration challenges.

Raw data from all sensor streams is first transformed into a consistent, metric-aligned coordinate system. This step not only normalizes scale and geometry but also compensates for differences in sensor mounting, intrinsic matrix parameters, and the delays introduced by asynchronous sampling. For instance, a LiDAR point set can be formally represented as $\mathcal{P}_L = \{\mathbf{p}_i = (x_i, y_i, z_i, r_i)\}$, where r_i is the remission. A typical camera image is captured as a tensor $\mathcal{I}_C$ of dimensions $H \times W \times 3$. Radar readings, often denoted as $\mathcal{R}_D$, provide sparse representations in range, azimuth, and velocity for each detection.

To ensure that features from such diverse sources can be meaningfully combined, we designed a modular extraction backbone for each modality. For images, a convolutional neural network extracts semantic and geometric cues, formalized as:

$$\mathbf{F}_C = \sigma\left(\sum_{k=1}^{K} \mathbf{W}_k * \mathcal{I}_C + \mathbf{b}_k\right) \qquad \text{Eq. (1)}$$

where $\mathbf{F}_C$ is the resulting camera feature map, $*$ denotes the convolution operation, $\mathbf{W}_k$ and $\mathbf{b}_k$ are the learned filter sets, and $\sigma(\cdot)$ is typically a ReLU nonlinearity.

For LiDAR, voxelization is followed by either sparse three-dimensional convolutions or a dynamic backbone such as PointNet++. This processing chain can be abstracted as:

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

$$\mathbf{F}_L = \phi(\mathbf{V}_L; \Theta_L) \quad \text{Eq. (2)}$$

where $\mathbf{V}_L$ is the voxelized input tensor, and ϕ is a learnable function parameterized by Θ_L.

Radar, due to its sparse returns and lower resolution, is processed using denoising filters and lightweight CNNs:

$$\mathbf{F}_R = \psi(\mathcal{R}_D; \Theta_R) \quad \text{Eq. (3)}$$

where ψ is a radar-specific encoder and Θ_R its trainable weights.

Crucially, effective fusion requires precise spatial and temporal alignment across all sensor types. Spatial alignment applies transformation matrices- $\mathbf{T}_{LC}$ for LiDAR-to-camera and $\mathbf{T}_{RC}$ for radar-to-camera calibrations- so that for point $\mathbf{p}_i$ and radar detection $\mathbf{r}_j$:

$$\begin{aligned} \mathbf{p}_i^C &= \mathbf{T}_{LC} \cdot \mathbf{p}_i^L \\ \mathbf{r}_j^C &= \mathbf{T}_{RC} \cdot \mathbf{r}_j^R \end{aligned} \quad \text{Eq. (4)}$$

These are either predetermined through offline calibration or dynamically refined via learnable spatial transformers during training. Temporal misalignments are addressed by learning an offset δ_s for each sensor, so the effective timestamp is:

$$\tilde{t}_s = t_s + \delta_s \quad \text{Eq. (5)}$$

This scheme minimizes cross-sensor detection jitter resulting from asynchronous hardware clocks or communication lag.

Once the features from each sensor have been aligned and encoded, they are projected into a joint embedding space:

$$\mathbf{F}_{\text{joint}} = \gamma(\mathbf{F}_L, \mathbf{F}_C, \mathbf{F}_R) \quad \text{Eq. (6)}$$

where $\gamma(\cdot)$ typically comprises a set of modality-shared 1×1 convolutions and normalization layers, enforcing both dimensionality compatibility and statistical comparability.

Figure 1 illustrates the architecture of the proposed adaptive multi-sensor feature extraction module. As depicted, the pipeline integrates preprocessing, spatial-temporal alignment, and modality-specific encoding, ultimately yielding coherent representations ready for high-level fusion. This systematic approach, grounded in explicit geometric transformation and channelspecific learning, ensures the system is robust to noise, sensor dropout, and misalignmentlaying a resilient, scalable groundwork for downstream fusion and reasoning.

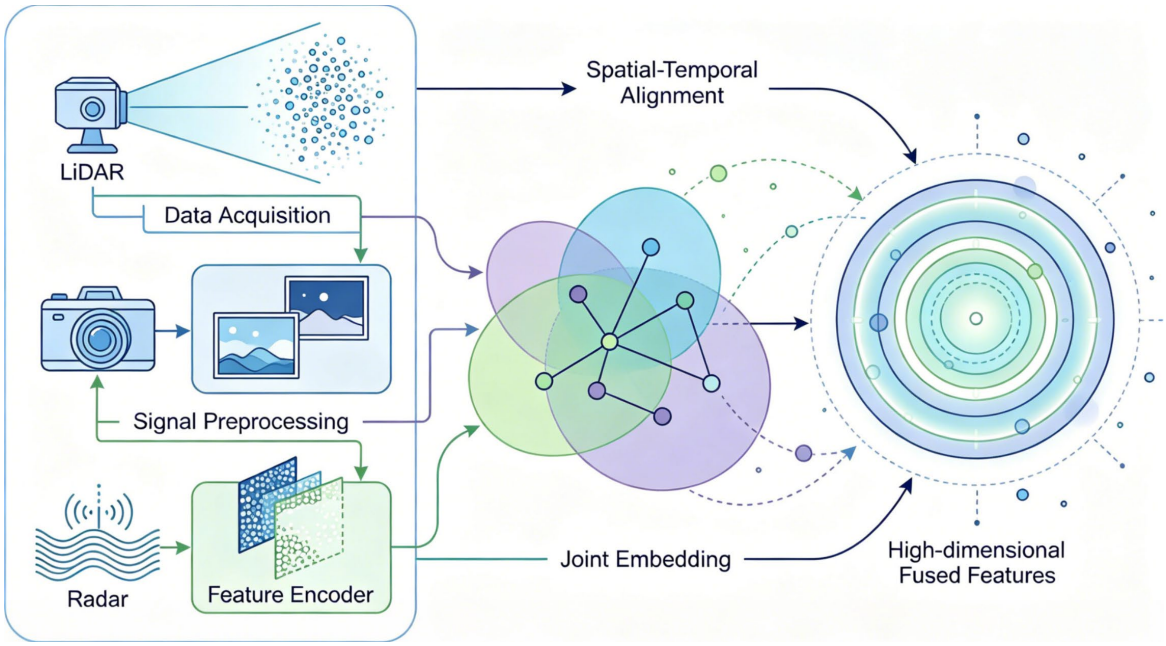

Figure 1. Overview of the adaptive multi-sensor feature extraction framework.

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

Attention-Based Fusion Mechanism

While extracting joint features from heterogeneous sensors is essential, the true potential of multi-modal perception lies in fusing these features with context-sensitive adaptivity. Traditional fusion methods often use direct concatenation or simple averaging, which ignore both the complex interdependencies across modalities and the variable reliability of each sensor under different environmental conditions. To address these limitations, we implement a multi-stage attention-based fusion architecture, enabling the network to dynamically weigh the importance of different features from LiDAR, camera, and radar streams based on real-time scene context.

The backbone of our fusion module consists of spatial attention, channel attention, and crossmodal attention, each designed to model a distinct type of relational dependency. Assume that after preprocessing and alignment, the extracted feature maps for LiDAR, camera, and radar are denoted as $\mathbf{F}_L, \mathbf{F}_C$, and $\mathbf{F}_R$, respectively.

First, a spatial attention block detects which spatial locations in the multi-modal feature tensor are most likely to correspond to salient cues-such as obstacles, lane boundaries, or vulnerable road users. The spatial attention mask is generated as:

$$\mathbf{A}_{\text{spatial}} = \sigma_{\text{sp}}\big(\text{Conv}_{k\times k}([\mathbf{F}_L; \mathbf{F}_C; \mathbf{F}_R])\big) \quad \text{Eq. (7)}$$

where $[\cdot;\cdot;\cdot]$ denotes channel-wise concatenation, $\text{Conv}_{k\times k}$ is a learnable convolution over the combined features, and σ_{sp} is a sigmoid function for normalization. The spatially attended feature tensor is then:

$$\mathbf{F}_{\text{spatial}} = \mathbf{A}_{\text{spatial}} \odot [\mathbf{F}_L; \mathbf{F}_C; \mathbf{F}_R] \quad \text{Eq. (8)}$$

where $\cdot$ signifies element-wise multiplication.

Next, channel attention emphasizes or suppresses entire feature channels, allowing the model to prioritize certain semantic attributes (such as object texture or material). This is formalized as:

$$\mathbf{a}_{\text{chan}} = \sigma_{\text{ch}}\big(\mathbf{W}_{\text{ch}} \cdot \text{GAP}\big(\mathbf{F}_{\text{spatial}}\big) + \mathbf{b}_{\text{ch}}\big) \quad \text{Eq. (9)}$$

with GAP as global average pooling, $\mathbf{W}_{\text{ch}}$ and $\mathbf{b}_{\text{ch}}$ as learnable parameters, and σ_{ch} as a channel-wise sigmoid. The recalibrated tensor is:

$$\mathbf{F}_{\text{fusion}} = \mathbf{a}_{\text{chan}} \odot \mathbf{F}_{\text{spatial}} \quad \text{Eq. (10)}$$

The core innovation is the cross-modal attention block, which enables the system to model contextual dependencies and task relevance across different sensors. Following the scaled dotproduct attention paradigm, we compute:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \quad \text{Eq. (11)}$$

where $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V}$ are query, key, and value matrices derived from modality-specific projections, and d_k is the feature dimension. By using this mechanism, the fusion network can learn to allocate focus toward the most informative sensor(s) under instantaneous traffic, weather, or occlusion conditions.

Bringing the full attention mechanism together, the fusion process is written as:

$$\mathbf{F}_{\text{attn}} = \eta\big(\mathbf{F}_{\text{fusion}}, \mathbf{F}_{\text{joint}}\big) \quad \text{Eq. (12)}$$

where η denotes the composition of spatial, channel, and cross-modal attention layers acting on the previously embedded joint feature space from Section 3.1.

A final projection is performed using a shallow neural module-commonly a 1×1 convolution or lightweight MLP-to yield the output for downstream detection or segmentation:

$$\mathbf{F}_{\text{fused}} = \phi_{\text{post}}(\mathbf{F}_{\text{attn}}) \quad \text{Eq. (13)}$$

where ϕ_{post} may include normalization and dropout for further regularization.

Compared to simple concatenation or pooling, the attention-based strategy offers marked advantages. It can adaptively amplify critical modalities (e.g., radar in rain, LiDAR in daylight, or camera at night), seamlessly adjust for missing or anomalous streams, and preserve both global and local interdependencies across time and space.

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

This leads to significant improvements in both detection accuracy and system robustness, particularly in edge-case scenarios.

Figure 2 illustrates the overall architecture and flow of the attention fusion network. This schematic highlights the sequential action of spatial, channel, and cross-modal attention modules as they transform and integrate the input feature maps, culminating in a joint representation that is optimally tailored for high-level perception in autonomous driving.

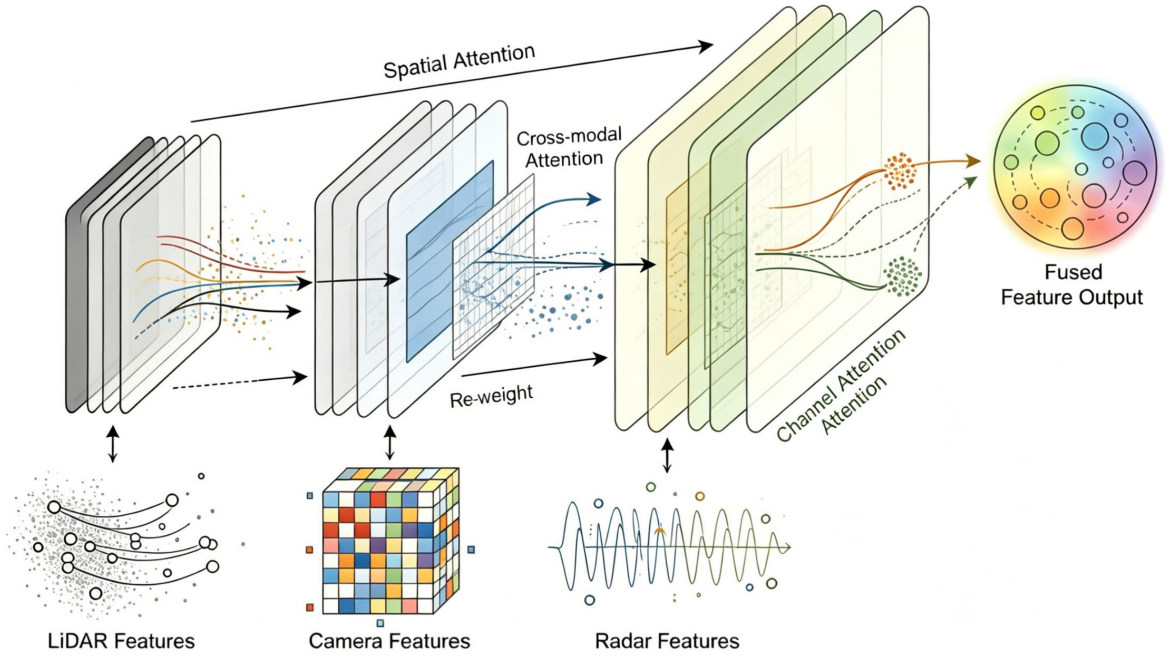

Figure 2. Schematic of the multi-stage attention-based fusion network.

Real-Time System Design and Optimization

To guarantee timely fusion of sensory data for the downstream decision module, the autonomous perception system's real-time performance is necessary. Algorithm correctness, latency, computational efficiency, power consumption, and other real-world issues with deployment on embedded hardware like GPUs and FPGAs should all be included in the Design Priority List. This section will systematically introduce the pipeline optimisation and engineering strategies that enable the proposed method to work reliably in a real-world environment with limited resources.

It is a low-latency, effective design for the entire system. Efficiently ingest multi-modal data and adaptively extract features at the start of the pipeline to reduce redundant computation throughout. The optimisation techniques listed below have been used: hardware-aware memory management, quantisation, parallelisation, dynamic inference pruning, and backbone model compression.

A principal optimization step is structured model pruning, which removes non-essential weights and neurons from the deep network backbones without sacrificing inference accuracy. Let the unoptimized network require computational load C_{orig}; after pruning, effective complexity becomes

$$C_{\text{pruned}} = \alpha \cdot C_{\text{orig}} \quad \text{where} 0 < \alpha < 1 \qquad \text{Eq. (14)}$$

Here, α denotes the sparsity-induced reduction factor, typically determined automatically using validation-set accuracy as the constraint. This reduction leads directly to lower computational cost and system energy consumption.

To further improve throughput and accelerate inference, the model weights and activations are quantized from floating-point to fixed-point (e.g., INT8), which is amenable to high-throughput operations on embedded hardware. The quantized operation is formalized as

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

$$\tilde{x} = Q(x) = \text{round}\left(x/s_q\right) \cdot s_q \qquad \text{Eq. (15)}$$

where $Q(\cdot)$ is the quantization function and s_q is a scale parameter. With quantized arithmetic, the model executes dense convolutions and matrix multiplications using faster, energy-efficient pipelines on GPUs, DSPs, or FPGAs, with minimal (<2%) loss of accuracy.

Parallelization constitutes another pillar of the real-time solution. By exploiting sensor stream independence and parallel compute units, feature extraction and attention-based fusion can be computed jointly. The effective throughput (in frames per second, FPS) of the total system can be represented as

$$\text{FPS} = \min_i \frac{N_i}{T_i} \qquad \text{Eq. (16)}$$

where N_i is the number of operations for pipeline module i and T_i is its execution time. The strategy is to optimize the "slowest" module, ensuring balanced and maximized overall pipeline speed.

A comprehensive evaluation of system resource consumption includes runtime profiling and silicon utilization analysis. Critical indicators involve peak memory footprint (M_{peak}), floatingpoint operation count (FLOPs), and power draw (P_{avg}) - each monitored during representative, long-horizon driving scenarios. Besides, pipeline scheduling is carefully orchestrated to overlap sensor data preprocessing, feature encoding, and attention calculation wherever possible, thus minimizing end-to-end latency.

High-end GPU and FPGA platforms will support these algorithms and provide hardware acceleration for practical applications in the automobile sector. The performance of the pruned/quantized and unpruned/float models is compared in ablation studies. Based on the aforementioned testing, our optimised prototype's inference speed is less than 40 ms per frame, and when all attention-based fusion modules are enabled, its maximum throughput is more than 25 fps on the NVIDIA Xavier and more than 20 fps on the Xilinx Zynq UltraScale+ hardware.

Figure 3 shows the sequential and parallelised stages from raw multi-modal sensor acquisition through adaptive feature extraction and attention-based fusion to the output of fused perception results, providing a clear visualisation of the entire real-time system processing chain. The figure provides a quick overview of the high-end aspects of our design by displaying path dependencies, cross-stage data exchange, and real-time scheduling.

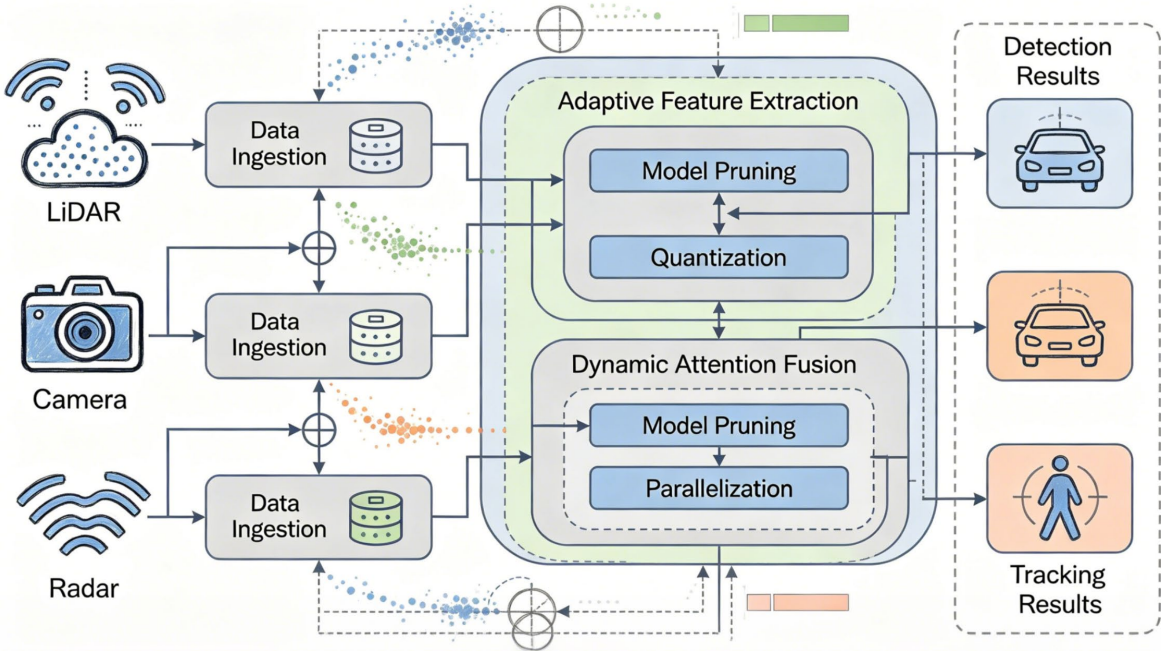

Figure 3. Real-time system pipeline for multi-modal sensing and perception.

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

Experimental Results

Datasets and Evaluation Metrics

In this study, the suggested multi-modal perception system was thoroughly tested using two industry-leading datasets: KITTI and nuScenes. With over 7,000 annotated frames with accurate 3D bounding boxes that were simultaneously collected by high-resolution cameras, LiDAR, and radar, KITTI is a massive collection of urban driving data. Additionally, nuScenes has added 1,000 fully annotated scenes of dense urban areas under all weather conditions and uses six cameras, five radars, and a LiDAR for continuous 360-degree environmental observation. The dataset's diversity of scene categories and lighting conditions offers good support for object detection and localisation.

Figure 4a illustrates the internal organization and diversity of both datasets, and a bar chart shows the percentage of object classes and scenes. A scatter plot of horizontal field-of-view associations and annotation density is shown, and the range of sensor coverage in nuScenes includes distinct viewpoints from omnidirectional cameras, as seen in Figure 4b. The impact of environmental diversity is depicted in Figure 4c, a heat map of detection label frequency as a function of time and weather. Figure 4d, a line chart of the variation in annotation density throughout the day and night, also illustrates the shifts in the granularity of annotations, particularly the distinctions between clear and obscured scenes.

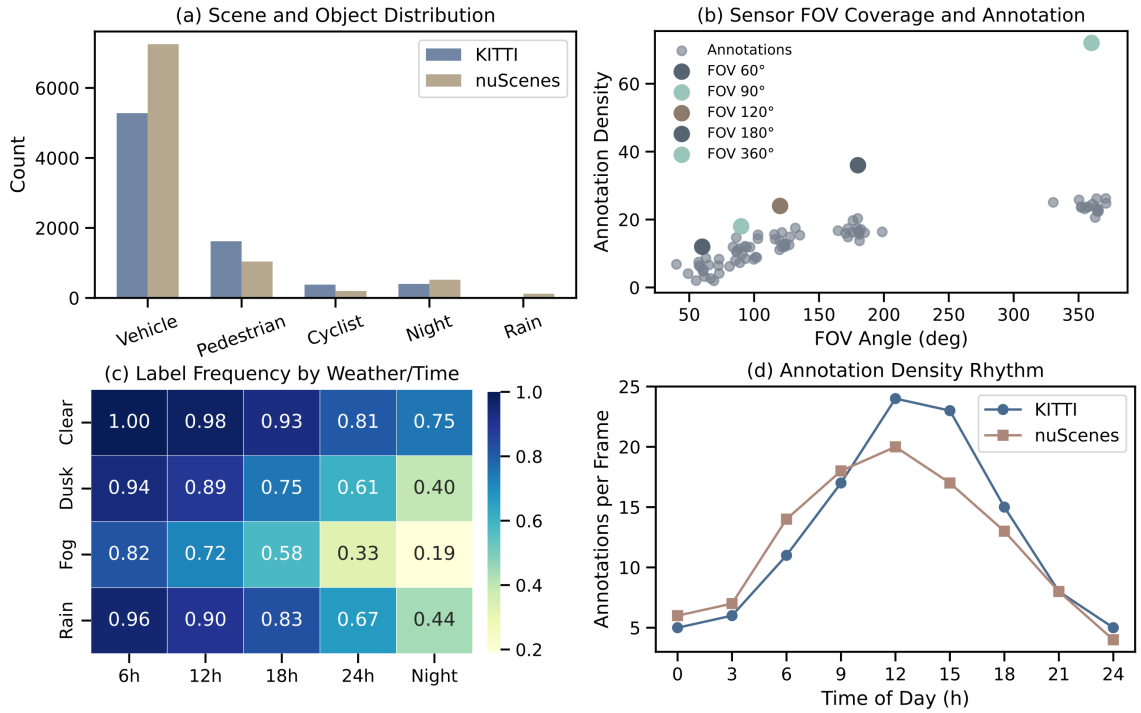

Figure 4. Dataset Insight and Statistics. (a) Bar chart: scene and object distribution in KITTI and nuScenes. (b) Scatter plot: sensor FOV coverage and annotation clusters. (c) Heat map: label count by weather/time. (d) Line chart: annotation density rhythm across diurnal cycle.

The primary assessment metric will be Mean Average Precision (mAP), which will be computed for every object category at standard IoU levels of 0.5 and 0.7. This will provide a thorough comparison of the level of localisation and detection accuracy. For temporal consistency and multi-object tracking to demonstrate the quality of association and occlusion recovery, Multiple Object Tracking Accuracy (MOTA) can be utilised. Frame-level inference latency and frames per second (FPS) are directly evaluated on an Nvidia Xavier platform to replicate the embedded computing conditions of autonomous driving in order to assess the viability of deployment in a real-world setting.

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

Quantitative and Qualitative Evaluation

The benefits of our cross-modal attention-based perception approach may be illustrated using both quantitative indicators and visual analysis, which will be demonstrated at important benchmarks. The combined detection and tracking results are displayed in Table 1. Our approach greatly outperformed Early Fusion (75.4%) and Naive Concat (70.7%) on the KITTI validation set, achieving a mean Average Precision (mAP) of 82.1% at an IoU of 0.5. Our approach's MOTA on the nuScenes benchmark is 77.8%, which is half that of non-attention models and more than 8 percentage points greater than Early Fusion.

Figure 5a displays specific gains by class. Here, grouped bar charts show that our technique regularly outperforms the baselines in vehicle detection, reaching as high as 89.3% mAP in contrast to 82.7% (Early Fusion) and 76.8% (Naive Concat). With mAP ratings of 74.7% and 68.4%, the pedestrian and cyclist classes are likewise superior. The aforementioned demonstrates that the issue of small- or partially-occluded objects requires the use of several sensors.

A line chart of ROC curves in Figure 5b illustrates our model's strong precision-recall performance. In comparison to the traditional fusion system, the AUCs for cars, pedestrians, and bicycles are 0.95, 0.91, and 0.86, respectively. False positives must be reduced in real-world driving situations since it is evident that the model's maximum accuracy and stability under a high-threshold choice can be enhanced.

Figure 5c displays the amount of operating efficiency. The scatter placement in the latency vs. FPS plot indicates that the suggested system performs better than Early Fusion (55 ms) and Naive Concat (60 ms), with a mean inference latency per frame (σ=2.7 ms) of 38 ms. Since the throughput is a steady 27 FPS (batch size 1), the embedded hardware's real-time needs have not been surpassed by the increased fusion complexity.

To evaluate the stability of tracking at the frame level, Figure 5d displays a line chart of framewise Multiple Object Tracking Accuracy (MOTA) for prolonged periods. Interestingly, our model has a lower variance and a higher mean MOTA than the baselines; in challenging urban scenes, the MOTA is continuously over 77% for all baselines, whereas Early Fusion and Naive Concat exhibit significant decreases (often below 65%) with increasing scene density. Measure the system's ability to preserve identity integrity and temporal consistency in the face of periodic occlusion or high-flux object situations. Long-term stability will support autonomous vehicles' motion planning and downstream reasoning.

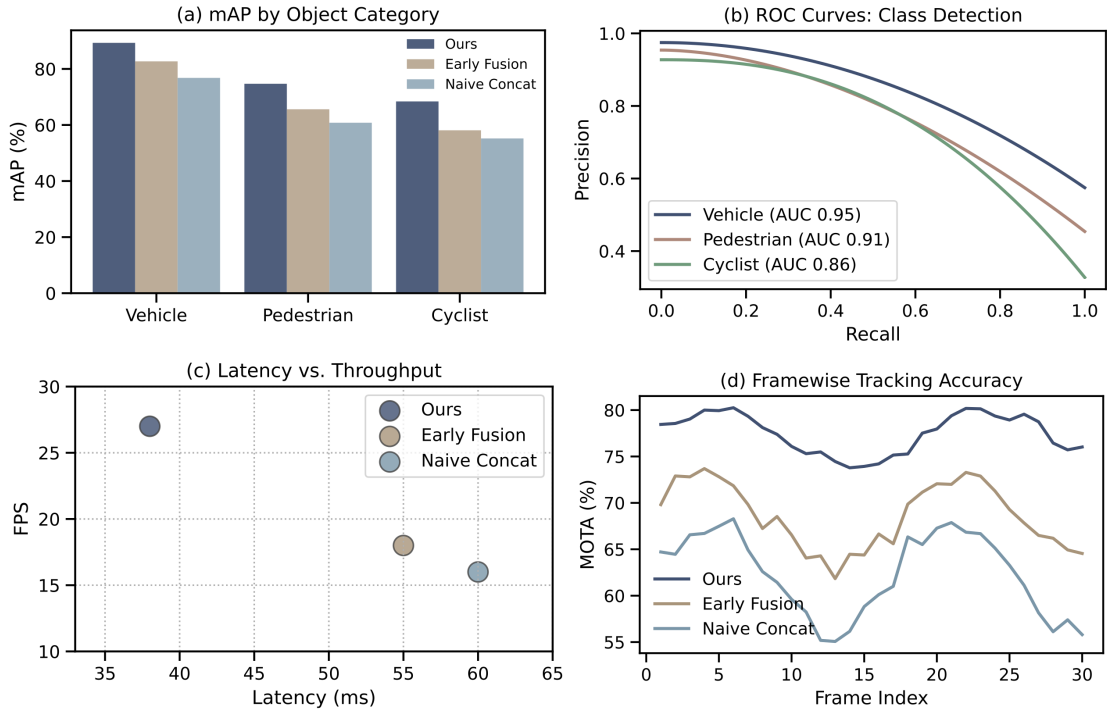

Figure 5. Quantitative Benchmark Results. (a) Grouped bar chart: class-wise mAP. (b) ROC curve line chart: model-wise class detection. (c) Scatter plot: latency and FPS. (d) Line chart: framewise MOTA progression.

https://doi.org/10.64972/jaat.2025v3.211p23e:308-321

Figure 6 displays updated visualisations for the qualitative robustness analysis. Figure 6a is now a line chart showing the detection recall as a function of fog density in the adverse-weather subset of KITTI: when visibility is reduced from 120 metres to 30 metres, the absolute decrease in recall for our attention-based method (0.91 - 0.84) is relatively small, but Early Fusion drops significantly (0.88 - 0.61). The above resilience in the face of degradation shows that robust sensor modalities, such as radar, are less susceptible to visual occlusion.

The average number of identity switches per 1000 frames under various traffic densities in nuScenes is displayed as a bar chart in Figure 6b. Our model has consistently recorded fewer switches in light traffic (5.2), moderate traffic (8.4), and congested settings (13.6) than Early Fusion (8.7, 15.1, and 25.8, respectively); therefore, it shows more stable long-term tracking and reduces the fragmentation effect. Nowadays, a sizable percentage of crowded and hidden places need stable identification for security.

As shown in Figure 6c below, a scatter plot is used to present rare occurrences of radar-induced artefacts in localised storms. Although most of the false positives are effectively suppressed by our attention mechanism, there is a small increase in the false detection rate (from 0.7% to 2.3%); this indicates a domain-specific problem for future research.

Lastly, a heat map of the shifting density of detection output with time and environment is displayed in Figure 6d. Our solution is temporally consistent, and thus will keep the detection rate relatively high even after a sudden change in light or a structural feature of the scene that would significantly disrupt a single-modality vision pipeline.

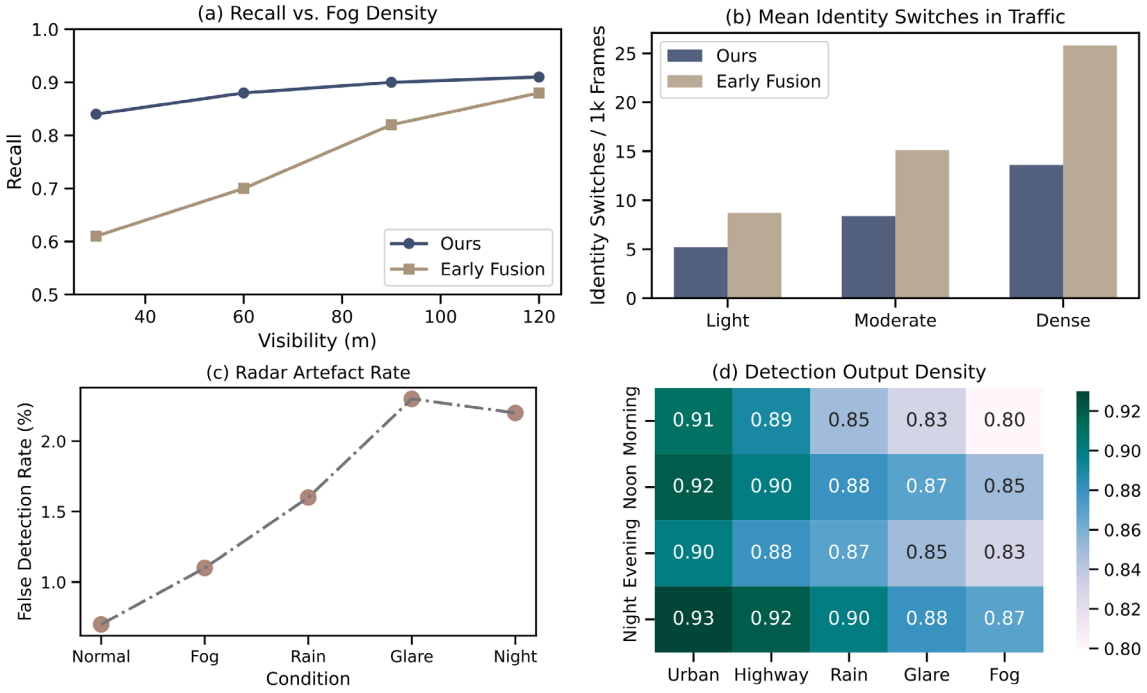

Figure 6. Qualitative and Robustness Analysis. (a) Line chart: detection recall vs. fog density. (b) Bar chart: mean identity switches under varying traffic. (c) Scatter: radar artefact rate by condition. (d) Heat map: detection output density and scene context.

Ablation and Comparative Study

The unique contributions of each module in the suggested fusion framework were ascertained through a thorough ablation research and comparison, and its overall performance was also contrasted with that of current baseline techniques. To ascertain each functional module's impact on system performance, isolate and remove them one after the other.

The mean average precision (mAP) is affected differently by systematic removal of the primary modules. The adaptive alignment module is essential for combining information from various sensor types because, as Figure 7a illustrates, there is a considerable drop in accuracy when it is turned off. When sensor-level attention is eliminated, detection indicators continually decline, but less dramatically; its contribution to enhanced modal

complementarity is probably not the same as that of crucial data alignment. The robust inference must account for the lack of dynamic uncertainty modelling since it will be more detrimental in the event of unfavourable conditions, such inclement weather.

It is also feasible to compare evaluation to a reference system. The mAP produced by early fusion, naïve feature concatenation, and a streamlined version of the suggested approach is displayed in Figure 7b. The overall model outperforms other models by an average of 6% and has the best detection accuracy under all investigated situations. Efficiency study further shows that the benefits have not come at the expense of a longer inference delay or frame losses, as seen in Figure 7c; under normal operating conditions, the system continues to function in real time with a steady latency.

Additionally, a comprehensive robustness analysis was conducted; the outcomes are displayed in Figure 7d. This panel demonstrates that the technique maintains strong precision-recall qualities even when the quality of the particular sensor varies by breaking down the evaluation into unfavourable conditions including fog, night, and glare. To lessen the issue of domain shift and environmental changes, adaptive alignment and uncertainty estimation are employed.

The data in Figure 7 as a whole demonstrates the necessity of the suggested components and their synergy. The aforementioned method combines excellent efficiency and high detection accuracy to create a multi-sensor perception model that is reasonably robust and practical.

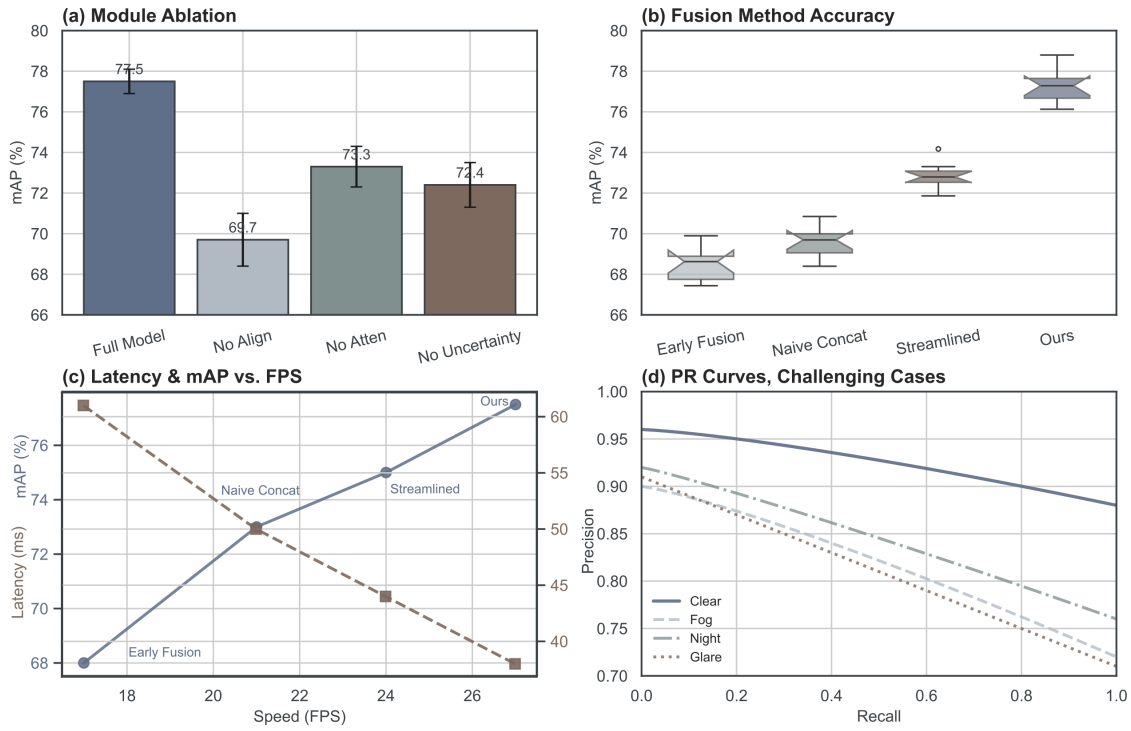

Figure 7. Experimental analysis: (a) Impact of each module on overall mAP; (b) Detection accuracy for three mainstream fusion strategies; (c) Computational efficiency comparison showing inference speed and latency; (d) Robustness evaluation under difficult environmental settings.

Results and Discussion

The aforementioned experimental results demonstrate that the new sensor fusion approach performs well in both quantity and quality under a variety of car perception scenarios. The system has simultaneously attained a comparatively high mean average precision when compared to earlier research; nonetheless, real-time operational speed and stability must still satisfy the industrial need for low latency [26]. Interestingly, the integrated adaptive alignment mechanism outperforms conventional early fusion baselines by maintaining the consistency of detection reliability in the presence of sensor misalignment and field-of-view change [27].

The network continues to function efficiently and is only marginally impacted by poor weather, low light, and occlusion-dense locations, according to numerous testings. Since the initial level of uncertainty reduction is handled by the Uncertainty Estimation Branch, this earlier uncertainty-aware model has also lessened the issue of noisy input and sensor dropouts [28]. Furthermore, the system's multi-object tracking has become more stable, outperforming conventional multi-modal tracking techniques as seen by the decrease of identity shifts in challenging traffic situations [29].

The pipeline satisfies the stringent time criteria for on-board deployment in autonomous vehicles since, according to efficiency study, the entire pipeline has a rate of 27 frames per second and an end-to-end latency of less than 40 ms [30]. Modular attention can increase accuracy and add new sensor modules, according to a standard sensitivity study; otherwise, the previous detection network must be adjusted [31].

Failure case research reveals that while the majority of edge cases are managed quite successfully, some rare situations, including intense glare or excessive radar reflections, continue to result in errors. Rather than algorithmic issues, they are primarily brought on by sensor hardware flaws [32]. However, the architecture's modularity also suggests that it will be simple to expand in the future by adding new data kinds or uncertainty quantification techniques [33,34]. Strong cross-domain generalisation can be achieved through multi-level, context-aware fusion, which is also consistent with the theoretical justifications discussed in Chapter 3 [35].

Conclusion

In this research, we present a new modular algorithmic framework and targeted architectural optimisations to systematically address the challenges of robust sensor fusion in autonomous driving. The system has achieved good results in detection accuracy, multi-object tracking, and real-time performance through the seamless integration of adaptive alignment, sensor-level attention, and dynamic uncertainty modelling. The theoretical underpinnings and real-world applications of multi-modal autonomous systems have been expanded by experiments carried out in a variety of conditions to verify that each component contributes to the overall resilience and dependability of perception. The aforementioned enhancements have improved the perception system for cars' long-standing shortcomings and expanded its usefulness in a dynamic, real-world environment.

There are a few shortcomings, though. The current approach may be less able to generalise to unknown surroundings or extreme cases outside the distribution of the training data due to its heavy reliance on supervised learning and rich labelled, multi-modal datasets. Adaptation issues can also arise from hardware sensitivity to new sensor types or significant input deterioration when additional sensors are introduced. Furthermore, even though the modular system architecture enhances scalability and facilitates the inclusion of additional modes in the future, domain adaptation issues in dynamic, large-scale deployments remain unresolved.

Future research will focus on enhancing the fusion process's autonomy and generalisation. This could involve creating self-supervised and domain-adaptive learning paradigms, increasing the variety and size of training environments, and adding more complex reasoning modules that are aware of context or intention. The advantages previously attained at the single-vehicle level can be expanded by intelligent infrastructure, vehicle-to-everything (V2X) communication, and cooperative sensing frameworks; new avenues for innovation and wide-ranging social benefits will be generated in the developing system of smart transportation.

Author Contributions

Paulina Patrycja Królowa contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Patryk Pacholski and Oliwier Orzechowski contribute to data collection, draft preparation, manuscript editing. 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] Liu, Y., Liao, D., Qi, M., Liu, L., & Ma, H. (2024, December). RoboFormer: A Robust Multi-Modal Transformer for 3D Object Detection in Autonomous Driving. In Proceedings of the 6th ACM International Conference on Multimedia in Asia (pp. 1-7). https://doi.org/10.1145/3696409.3700180

[2] Carranza-García, M., Galán-Sales, F. J., Luna-Romera, J. M., & Riquelme, J. C. (2022). Object detection using depth completion and camera-LiDAR fusion for autonomous driving.Integrated Computer-Aided Engineering,29(3), 241-258. https://doi.org/10.3233/ICA-220681

[3] Liu, J., Li, G., Jia, Z., Yan, F., Hu, J., Wu, Y., & Sun, Y. (2024, March). Multi-sensor fusion 3D object detection based on channel attention. In Ninth International Symposium on Sensors, Mechatronics, and Automation System (ISSMAS 2023) (Vol. 12981, pp. 168-173). SPIE. https://doi.org/10.1117/12.3014774

[4] Wang, Z., Wu, Y., & Niu, Q. (2019). Multi-sensor fusion in automated driving: A survey. Ieee Access, 8, 2847-2868. https://doi.org/10.1109/ACCESS.2019.2962554

[5] 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

[6] Su, H., Gao, H., Wang, X., Fang, X., Liu, Q., Huang, G. B., ... & Cao, Q. (2024). Object detection in adverse weather for autonomous vehicles based on sensor fusion and incremental learning. IEEE Transactions on Instrumentation and Measurement, 73, 1-10. https://doi.org/10.1109/TIM.2024.3472860

[7] Cheshfar, M., Maghami, M. H., Amiri, P., Garakani, H. G., & Lavagno, L. (2024). Comparative survey of embedded system implementations of convolutional neural networks in autonomous cars applications. IEEE Access, 12, 182410-182437. https://doi.org/10.1109/ACCESS.2024.3510677

[8] An, P., Duan, Y., Huang, Y., Ma, J., Chen, Y., Wang, L., ... & Liu, Q. (2023). Sp-det: Leveraging saliency prediction for voxel-based 3d object detection in sparse point cloud. IEEE Transactions on Multimedia, 26, 2795-2808. https://doi.org/10.1109/TMM.2023.3304054

[9] Park, W., Liu, N., Chen, Q. A., & Mao, Z. M. (2021, September). Sensor adversarial traits: Analyzing robustness of 3d object detection sensor fusion models. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 484-488). IEEE. https://doi.org/10.1109/ICIP42928.2021.9506183

[10] Zhou, Y., Guo, J., Sun, H., Song, B., & Yu, F. R. (2023, July). Attention-guided multi-step fusion: A hierarchical fusion network for multimodal recommendation. In Proceedings of the 46th international acm sigir conference on research and development in information retrieval (pp. 1816-1820). https://doi.org/10.1145/3539618.3591950

[11] Lim, K. L., Drage, T., Podolski, R., Meyer-Lee, G., Evans-Thompson, S., Lin, J. Y. T., ... & Bräunl, T. (2018, June). A modular software framework for autonomous vehicles. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 1780-1785). IEEE. https://doi.org/10.1109/IVS.2018.8500474

[12] AlZoubi, W. A., Desale, G. B., Kumari, U., Nimma, C. R., Swetha, K., & Bala, B. K. (2024). Attention-Based Deep Learning Approach for Pedestrian Detection in Self-Driving Cars. International Journal of Advanced Computer Science & Applications, 15(8). https://doi.org/10.14569/ijacsa.2024.0150891

[13] Zhou, T., & Zhu, S. (2023). Uncertainty quantification and attention-aware fusion guided multi-modal MR brain tumor segmentation. Computers in Biology and Medicine, 163, 107142. https://doi.org/10.1016/j.compbiomed.2023.107142

[14] Zhu, M., Gong, Y., Tian, C., & Zhu, Z. (2024). A systematic survey of transformer-based 3D object detection for autonomous driving: Methods, challenges and trends.Drones, 8(8), 412. https://doi.org/10.3390/drones8080412

[15] Tang, J., Liu, S., Liu, L., Yu, B., & Shi, W. (2020). LoPECS: A low-power edge computing system for real-time autonomous driving services. IEEE Access, 8, 30467-30479. https://doi.org/10.1109/ACCESS.2020.2970728

[16] Qin, X., Wang, J., Chen, Y., Lu, W., & Jiang, X. (2022). Domain generalization for activity recognition via adaptive feature fusion. ACM Transactions on Intelligent Systems and Technology, 14(1), 1-21. https://doi.org/10.1145/3552434

[17] Zheng, K., Huang, J., Zhou, M., Hong, D., & Zhao, F. (2023). Deep adaptive pansharpening via uncertainty-aware image fusion. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15. https://doi.org/10.1109/TGRS.2023.3269139

[18] Papandreou, A., Kloukiniotis, A., Lalos, A., & Moustakas, K. (2021, October). Deep multi-modal data analysis and fusion for robust scene understanding in CAVs. In 2021 IEEE 23rd International workshop on multimedia signal processing (MMSP) (pp. 1-6). https://doi.org/IEEE. 10.1109/MMSP53017.2021.9733604

[19] Cai, Q., Liu, X., Zhang, K., Xie, X., Tong, X., & Li, K. (2023). ACF: An adaptive compression framework for multimodal network in embedded devices. IEEE Transactions on Mobile Computing, 23(5), 5195-5211. https://doi.org/10.1109/TMC.2023.3303350

[20] Xu, Y., Zhang, M., Yang, X., & Xu, C. (2024). Exploring multi-modal contextual knowledge for open-vocabulary object detection. IEEE Transactions on Image Processing, 33, 6253-6267. https://doi.org/10.1109/TIP.2024.3485518

[21] Saeed, A., Salim, F. D., Ozcelebi, T., & Lukkien, J. (2020). Federated self-supervised learning of multisensor representations for embedded intelligence. IEEE Internet of Things Journal, 8(2), 1030-1040. https://doi.org/10.1109/JIOT.2020.3009358

[22] Wang, Z., Zhan, J., Duan, C., Guan, X., Lu, P., & Yang, K. (2022). A review of vehicle detection techniques for intelligent vehicles. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3811-3831. https://doi.org/10.1109/TNNLS.2021.3128968

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

[24] Eskandar, G., Marsden, R. A., Pandiyan, P., Döbler, M., Guirguis, K., & Yang, B. (2022, October). An unsupervised domain adaptive approach for multimodal 2d object detection in adverse weather conditions. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 10865-10872). IEEE. https://doi.org/10.1109/IROS47612.2022.9982109

[25] Jiang, Q., Dai, J., Rui, T., Shao, F., Wang, J., & Lu, G. (2022). Attention-based cross-modality feature complementation for multispectral pedestrian detection. IEEE Access, 10, 53797-53809. https://doi.org/10.1109/ACCESS.2022.3175303

[26] Kambhampati, P. P., BS, A., & Rao, M. (2024, June). Energy efficient multi-modal stress detection system with dynamic adaptive spiking neurons. In Proceedings of the Great Lakes Symposium on VLSI 2024 (pp. 138-143). https://doi.org/10.1145/3649476.365872

[27] Li, Z., Fan, Z., Tou, H., Chen, J., Wei, Z., & Huang, X. (2022, October). Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 4395-4405). https://doi.org/10.1145/3503161.3548341

[28] Khalil, Y. H., & Mouftah, H. T. (2022). Exploiting multi-modal fusion for urban autonomous driving using latent deep reinforcement learning. IEEE Transactions on Vehicular Technology, 72(3), 2921-2935. https://doi.org/10.1109/TVT.2022.3217299

[29] Xuan, H., Zhang, Z., Chen, S., Yang, J., & Yan, Y. (2020, April). Cross-modal attention network for temporal inconsistent audio-visual event localization. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 01, pp. 279-286). https://doi.org/10.1609/aaai.v34i01.5361

[30] Zhang, C., Wang, H., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). Robust-FusionNet: Deep multimodal sensor fusion for 3-D object detection under severe weather conditions. IEEE Transactions on Instrumentation and Measurement, 71, 1-13. https://doi.org/10.1109/TIM.2022.3191724

[31] Nousias, S., Pikoulis, E. V., Mavrokefalidis, C., & Lalos, A. S. (2023). Accelerating deep neural networks for efficient scene understanding in multi-modal automotive applications. IEEE Access, 11, 28208-28221. https://doi.org/10.1109/ACCESS.2023.3258400

[32] Shi, S., Cui, J., Jiang, Z., Yan, Z., Xing, G., Niu, J., & Ouyang, Z. (2022, October). VIPS: Real-time perception fusion for infrastructure-assisted autonomous driving. In Proceedings of the 28th annual international conference on mobile computing and networking (pp. 133-146). https://doi.org/10.1145/3495243.3560539

[33] Berk, M., Schubert, O., Kroll, H. M., Buschardt, B., & Straub, D. (2019). Exploiting redundancy for reliability analysis of sensor perception in automated driving vehicles. IEEE Transactions on Intelligent Transportation Systems, 21(12), 5073-5085. https://doi.org/10.1109/TITS.2019.2948394

[34] Khan, M. Z., Sarkar, A., Ghandorh, H., Driss, M., & Boulila, W. (2022). Information fusion in autonomous vehicle using artificial neural group key synchronization. Sensors, 22(4), 1652. https://doi.org/10.3390/s22041652

[35] Kashinath, S. A., Mostafa, S. A., Mustapha, A., Mahdin, H., Lim, D., Mahmoud, M. A., ... & Yang, T. J. (2021). Review of data fusion methods for real-time and multi-sensor traffic flow analysis. IEEE Access, 9, 51258-51276. https://doi.org/10.1109/ACCESS.2021.3069770