Journal of Applied Automation Technologies

Research Paper

Multi-Sensor Fusion Algorithms for Autonomous Vehicles

Violetta Mikołajczyk[1,*] and Hanna Płocharska[1]

[1] Faculty of Mechatronics and Automatics, Lodz University of Technology, Lodz, 90-924, Poland
*Corresponding author: violetta.miko@p.lodz.pl

Abstract. By using sensor fusion and deep learning technologies, the perception capabilities of autonomous driving systems are advancing. This paper proposes an all-terrain, all-weather, single-sensor fusion framework. The four components of the system are cameras, LiDAR, radar, and inertial/ultrasonic sensors. The neural network architecture precisely integrates the extracted features in both time and space through attention-based fusion. A large number of experiments on distributed high-performance computing systems were conducted on the KITTI and nuScenes datasets. According to the above experiments, the proposed hybrid fusion model outperforms early fusion and single-modal baseline methods, achieving an accuracy of 0.91 during the day and an accuracy of 0.80 at night or under adverse weather conditions. The range of the median and the standard deviation of the results are both very small under different weather conditions. Based on the ablation experiments, each sensor has its own advantages. In terms of system efficiency, they can reach up to 27 frames per second on the RTX 3090 graphics card, and all have stable power consumption and memory. In summary, the aforementioned structure strikes an appropriate balance between detection performance, computational efficiency, and robustness to sensor degradation. This lays the foundation for the application of autonomous driving in real life.

Keywords: *Autonomous Vehicles, Object Detection, Real-Time Processing, Multi-Modal Integration, Embedded Systems*

Received on 24 December 2024, Accepted on 29 May 2025, Published on 09 June 2025

Introduction

With the development of autonomous driving technology, the demand for perception systems to accurately identify various dynamic and unstructured objects in the road environment is increasing [1]. In order to ensure the safety of autonomous vehicles, a robust perception system is required. Inertial measurement units, LiDAR, radar, and cameras are typically components of this system [2]. The first-generation algorithms may be affected by adverse weather, low light conditions, and urban populations [3]. In order to enhance environmental perception and leverage the advantages of multiple sensors, multi-sensor fusion technology has recently begun to be adopted [4]. Early probabilistic frameworks help integrate important data [5]. Based on deep learning technology, it has recently been used for the fusion and improved extraction of heterogeneous data [6]. The aforementioned new model has expanded the operational range of autonomous driving and improved scene recognition and grouping capabilities [7]. As new problems arise in practice, many now believe that traditional algorithms can no longer solve these issues and that new algorithms are needed [8].

When deploying multi-sensor fusion technology for autonomous driving in the real world, there are still some serious issues [9]. Sensor synchronization errors, calibration drift, and mismatched data often lead to a decrease in the system's actual detection rate and speed [10]. Decision-level, feature-level, and data-level fusion strategies will have different impacts on the overall stability of the system, as well as its resistance to certain sensor failures and environmental changes [11]. Deep fusion networks are generally more accurate, but they are more expensive and less suitable for low-latency and resource-constrained embedded systems [12]. Mean Average Precision (mAP) and other traditional evaluation metrics have failed to address the issues that arise in

high-risk and rare applications [13]. There is no public dataset for testing general fusion algorithms that includes variations in light and shadow, adverse weather conditions, or partial occlusions [14]. In light of the aforementioned shortcomings, it is imperative to immediately establish a unified and comprehensive evaluation system to ensure the reliability of algorithms in all operational driving environments [15].

This paper proposes a unified multi-sensor fusion framework to achieve robustness in autonomous driving. Here are the precedents. First, a fusion architecture is proposed. This architecture is designed to handle the differences and misalignments of various sensors and ensure stable operation under various environmental and operational uncertainties. Secondly, establish a general experimental process that covers various lighting and weather conditions as well as occlusion scenarios, and comprehensively test the fusion strategies. Thirdly, ablation studies and efficiency analysis will be conducted. The purpose of this paper is to provide specific references on the trade-offs between computational cost, robustness, and perception accuracy. The aforementioned advancements lay the foundation for the deployment and further development of autonomous driving systems with flexibility and real-time capabilities.

Related Work

Traditional Multi-Sensor Fusion Methods

Robots and autonomous driving systems initially used rule-based and probabilistic methods to address the shortcomings of single-sensor systems in multi-sensor fusion [16]. At the sensor level, early fusion methods used Kalman filtering, Bayesian inference, and Dempster-Shafer theory to combine data [17]. Although data-level fusion can reduce noise and stabilize signals, it is usually not applicable to handling fundamentally asynchronous or different data from different sensors [18]. Joint histograms and feature vector concatenation are relatively new feature-level methods. They improve robustness by combining high-level representations from independent sensor channels [19]. Decision-level fusion makes the final decision by using the outputs of all sensors through a weighted confidence model or voting scheme [20]. The aforementioned traditional methods are mathematically simple and effective in structured environments, but they are usually not reliable enough under high dynamic conditions, when encountering sensor failures, or in the non-linear scenarios of modern autonomous driving.

Deep Learning-Based Fusion Algorithms

Deep learning has recently been used to integrate various data, and it can help automatically learn the best feature representations and their interconnections [21]. Early fusion networks are typically directly connected to a single neural network that processes low-level feature extraction or raw sensor data [22]. According to the noise or partial degradation characteristics of each sensor, different weights are assigned in multi-branch or attention-based models to enhance the robustness of the model [23]. Data, feature, or decision-level can be used. Due to its flexibility and ability to learn higher-order correlations between different modalities, feature-level fusion is becoming increasingly popular [24]. These deep models have already reached the limits of detection and segmentation accuracy; they typically require a large amount of data and computational resources to achieve good generalization.

Public Datasets and Evaluation Metrics

Several large-scale public datasets have already been created for research on multi-sensor fusion in autonomous driving, to promote progress and provide reliable benchmarks. Many popular datasets, such as KITTI, nuScenes, and the Waymo Open Dataset, have been collected in various environments under different conditions, providing rich real-world annotations for all these domains. To promote the development of fusion algorithms, there are datasets of varying difficulty, such as simple highway scenes and complex urban intersections. These evaluation metrics are common: precision, recall, mean Average Precision (mAP), and F1-score. These metrics showcase different aspects of detection and classification capabilities. Due to the higher demands on vehicle reliability in unusual driving conditions, some researchers have recently begun to focus more on robustness metrics and edge case testing in their work. Overall, these shared resources and standardized procedures lay the foundation for fair comparisons and reproducible experiments in the field.

https://doi.org/10.64972/jaat.2025v3.209p22e:294-307

Methodology

Sensor Data Preprocessing

High-quality multi-sensor fusion perception requires basic preprocessing of sensor data. Due to hardware limitations of sensors, network delays, and other factors, time and spatial inconsistencies often occur in autonomous driving. Therefore, synchronization is also necessary at this time. Consider the sensors $\mathcal{S} = \{S_1, \dots, S_n\}$, which collect environmental data at times $\mathbf{t} = \{t_1, \dots, t_n\}$. In order to temporally align each observation, alignment offsets $\tau = \{\tau_1, \dots, \tau_n\}$ are introduced to ensure they are consistent with each other. Therefore, regularization optimization strategies can be used to determine the minimum time offset that is both accurate and stable for the system:

$$\underset{\tau}{argmin} \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left((t_i + \tau_i) - \left(t_j + \tau_j \right) \right)^2 + \lambda_\tau \|\tau\|^2 \quad \text{Eq.(1)}$$

Once temporal alignment is addressed, spatial alignment through extrinsic calibration becomes equally critical. Each sensor's local coordinate frame must be mapped to a shared world reference. This transformation for any source point $\mathbf{p}_{\text{source}}$ involves a rotation $\mathbf{R}$ and translation $\mathbf{T}$, often computed via calibration targets or geometric approaches:

$$\boldsymbol{p}_{target} = \boldsymbol{R} \cdot \boldsymbol{p}_{source} + \boldsymbol{T} \quad \text{Eq.(2)}$$

The average reprojection error represents the calibration quality of heterogeneous sensors. This metric represents the difference between the actual position of the measurement point and the position determined by the calibration parameters. This difference is referred to as systematic alignment error:

$$E_{reproj} = \frac{1}{M} \sum_{m=1}^{M} \left\| \boldsymbol{p}_{m,\, measured} - \left(\boldsymbol{R} \cdot \boldsymbol{p}_{m,\, ref} + \boldsymbol{T} \right) \right\|_2 \quad \text{Eq.(3)}$$

Beyond alignment, noise and artifacts from environmental uncertainties or sensor defects must be mitigated. Denoising operators $\mathcal{D}(\cdot)$ are tailored to each modality. For example, statistical outlier removal in LiDAR can be realized as:

$$\tilde{x}_i = \begin{cases} x_i, & if |x_i - \mu| < k\sigma \\ discard, & otherwise \end{cases} \quad \text{Eq.(4)}$$

where μ and σ are the local mean and deviation, ensuring retention of valid points while filtering outliers robustly.

Finally, data augmentation is essential in supporting model generalization in rare or critical driving scenarios. Each clean data instance undergoes multiple stochastic transforms $\mathcal{T}_k$ – including rotations, intensity perturbations, or synthetic occlusions - increasing both the size and diversity of the training dataset:

$$\mathcal{X}_{aug} = \bigcup_{k=1}^{K} \mathcal{T}_k(\mathcal{X}_{clean}) \quad \text{Eq.(5)}$$

This practice ensures the fusion model does not merely memorize but learns representations robust to the complex, often adversarial nature of real-world environments.

Fusion Network Architecture Design

An effective multi-sensor perception fusion network should be able to flexibly display and integrate these sources. As shown in Figure 1, the proposed architecture designs independent backbone encoders for multiple modalities. These modalities include convolutional neural networks (CNN) for camera images, point-based transformers for LiDAR data, and temporal convolutional layers for radar sequences. This approach enables the model to extract rich and modality-specific features, resulting in intermediate representations $\mathbf{F}_c, \mathbf{F}_l$ and $\mathbf{F}_r$ computed in parallel for each data stream.

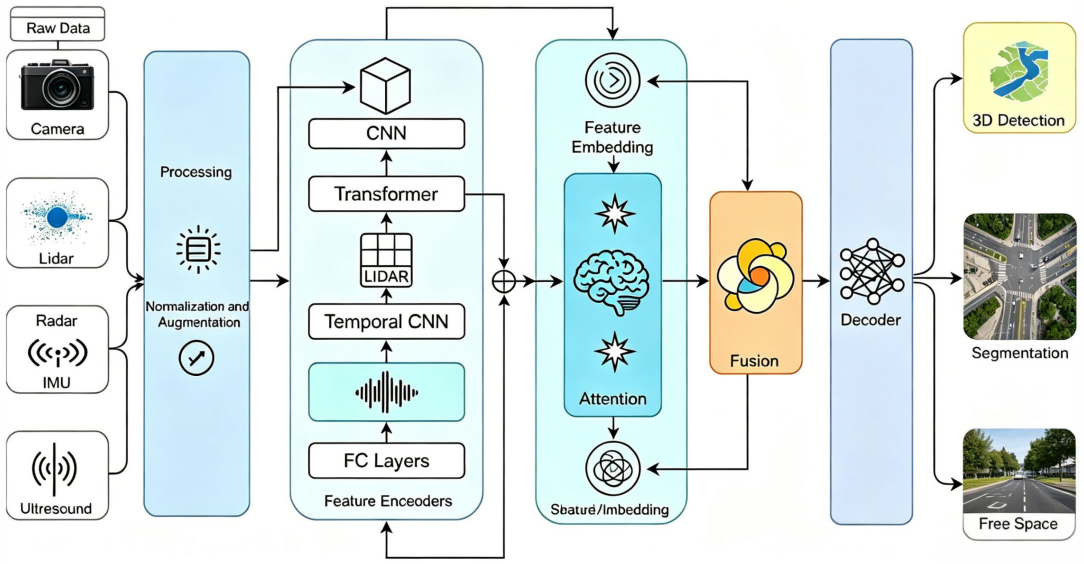

Figure 1. High-level architecture of the proposed multi-sensor fusion neural network

Before these attributes can be used for subsequent calculations, they must be mapped to the same semantic space. Apply a learned linear transformation to the feature maps of each modality, and then consistently fuse them for subsequent processing:

$$\boldsymbol{z}_s = \boldsymbol{W}_s \boldsymbol{F}_s + \boldsymbol{b}_s \quad \text{Eq.(6)}$$

where $\mathbf{W}_s$ and $\mathbf{b}_s$ are parameters learned independently for each modality s, and $\mathbf{z}_s$ denotes the embedded feature vector. This step ensures that differences in feature dimensionality or statistical distribution across sensors do not hinder holistic integration.

To combine these aligned features, a cross-modal attention module is applied. This mechanism enables the network to focus on the most informative and reliable sources in varying scenarios. The attention logit for assessing the relationship between modalities i and j is given by:

$$e_{ij} = \boldsymbol{q}_i^{\top} \boldsymbol{k}_j + b_{ij} \quad \text{Eq.(7)}$$

where $\mathbf{q}_i$ and $\mathbf{k}_j$ represent query and key vectors obtained from learned projections of the modality embeddings, and b_{ij} is a bias term. These logits encode the mutual compatibility and relevance between features extracted from different sensors.

To guarantee that fusion weights capture relative importance, the logits are normalized through a softmax operation across all available sensor branches for each context position:

$$\alpha_{ij} = \frac{exp\,(e_{ij})}{\sum_{k=1}^{n} exp\,(e_{ik})} \quad \text{Eq.(8)}$$

This normalization produces a set of attention weights α_{ij} that sum to one and allow dynamic prioritization. In practice, the network may assign higher weights to LiDAR features at night or elevate radar when vision is occluded.

With attention weights established, the final fused feature is calculated as a weighted sum of each modality's embedding:

$$\boldsymbol{z}_{fused} = \sum_{s=1}^{n} \alpha_s \boldsymbol{z}_s \quad \text{Eq.(9)}$$

Here, α_s encapsulates the contextual trust in sensor s, ensuring the output feature is robust to sensor failures and environmental uncertainty. Subsequently, this fused representation is used as input for the shared decoder network. There are multiple output heads responsible for performing various tasks, including semantic segmentation, 3D object detection, and free space estimation.

At the same time, the network is optimized using a combined loss function to balance detection, segmentation, and other auxiliary learning objectives, achieving stability and generalization capabilities:

$$L_{total} = \lambda_{det} \cdot L_{det} + \lambda_{seg} \cdot L_{seg} + \lambda_{aux} \cdot L_{aux} \quad \text{Eq.(10)}$$

The loss coefficients λ_{det}, λ_{seg}, and λ_{aux} are tuned to balance the network's emphasis according to deployment priorities, such as maximizing precision versus computational speed.

This modular and attention-based design can reliably handle missing or noisy sensor data. Moreover, various components such as task-specific decoders, sensor arrangements, and backbone networks can be easily replaced to quickly adapt and conduct ablation studies.

Fusion Workflow and Implementation

For industry standards of early, late, and hybrid fusion, the workflow of multi-sensor fusion is systematic and incremental, and sufficiently flexible. Early fusion allows the raw or lightly preprocessed sensor streams to be concatenated before feature extraction.

$$\boldsymbol{x}_{early} = [\mathcal{N}_1(x_1) \| \mathcal{N}_2(x_2) \| \ldots \| \mathcal{N}_n(x_n)] \quad \text{Eq.(11)}$$

where $\mathcal{N}_i$ is any signal normalization and all sensors are projected to a shared spatial/temporal domain.

In late fusion, feature extraction and inference are performed separately for each sensor, and only the results (e.g., predicted labels or bounding boxes) are combined at the decision level.

$$y_{final} = Fuse([y_1, y_2, \ldots, y_n]) \quad \text{Eq.(12)}$$

Here, the fusion operation may be a learned weighted sum, a confidence-weighted voting scheme, or another task-suited aggregator.

Hybrid fusion employs learnable coupling at strategic network layers, capturing richer crossmodal relationships. Intermediate representations from each modality at layer l, i.e., $\mathbf{h}_i^{(l)}$, are transformed together with:

$$\boldsymbol{h}_{fused}^{(l)} = FusionLayer^{(l)}([\boldsymbol{h}_1^{(l)}, \boldsymbol{h}_2^{(l)}, \ldots, \boldsymbol{h}_n^{(l)}]) \quad \text{Eq.(13)}$$

To implement this theory, a modular block-based pipeline was established for data collection and labeling; preprocessing for synchronization and denoising; feature extraction based on a backbone network; decision computation; adaptive fusion at multiple fusion points; and inference output. As shown in Figure 2, the pipeline is suitable for large-scale, high-throughput, and memory-efficient design. The total processing time of the main modules for batch inference determines the system-level batch inference throughput:

$$Throughput_{batch} = \frac{N_{sample}}{T_{prep} + T_{forward} + T_{fuse}} \quad \text{Eq.(14)}$$

where N_{sample} is the batch size, and T_{prep}, T_{forward}, T_{fuse} are the preprocessing, forward inference, and fusion latencies, respectively.

The aforementioned design concept will build a universal platform to facilitate relatively simple benchmarking and widespread use in many areas of sensor systems and autonomous vehicles.

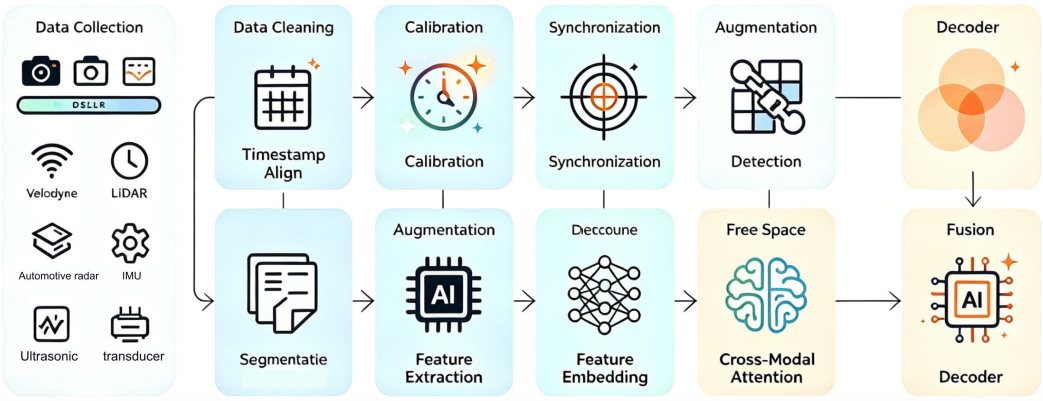

Figure 2. Block diagram of the end-to-end sensor fusion workflow.

https://doi.org/10.64972/jaat.2025v3.209p22e:294-307

Experiments and Analysis

Experimental Setup

All experiments in this paper were conducted on a high-performance distributed computing platform. The platform includes 512 GB of RAM, an AMD EPYC 7742 64-core CPU, and four NVIDIA RTX 4090 GPUs, each with 24 GB of dedicated memory. Using CUDA 12.1 and cuDNN 8.9 for GPU acceleration in deep learning, with corresponding optimizations, and Ubuntu 22.04 LTS as the software foundation. PyTorch 2.1 and Python 3.10 are the foundational libraries for the experimental code. To reduce computational and memory requirements, mixed precision training and the AdamW optimizer were chosen. Distributed synchronous SGD and multi-node scaling maintain statistical consistency for large-batch training across all runs [25].

During evaluation, KITTI and nuScenes are two reliable and distinct benchmarks. A total of 6,432 training scenes, 1,380 validation scenes, and 1,380 test scenes were obtained, with fully synchronized image, point cloud, and radar frame sequences selected solely from KITTI, and a 70/15/15 split at the scene level was performed to prevent data leakage. The nuScenes dataset was processed in the same way, containing only fully aligned multi-sensor segments. In the end, nuScenes obtained 19,926 training samples, 4,271 validation samples, and 4,271 test samples. To ensure that operational diversity reflects real-world deployment scenarios, the two datasets were stratified by scene, weather, and urban-rural ratio [26].

Robustness in preprocessing was ensured by enforcing standardized input for each modality, as described in Section 3. For KITTI, image data were resized to 1242×375 pixels and nuScenes to 1600×900. LiDAR point clouds were discretized into voxel grids of $0.1\ \mathrm{m}^3$, and radar data were normalized to 128×128 range-intensity maps. All samples underwent precise spatiotemporal synchronization, with extrinsic calibration errors verified to remain consistently below a centimeter. Augmentation protocols included random rotations up to $\pm 10^{\circ}$, multiplicative intensity jitter between 0.9 and 1.1, and synthetic simulation of adverse weather conditions (such as periodic fog or synthetic precipitation), thereby augmenting the dataset's ability to challenge model generalizability [27].

Networks were initialized using the Kaiming scheme, chosen for its empirical stability in deep architectures. Training proceeded for 80 epochs with a batch size of 16 per GPU. The learning rate started at 0.002 and decayed to 10^{-5} via cosine annealing. In the absence of improvement in validation loss, early stopping was used, with a dropout rate of 0.2 and L2 regularization ($\lambda = 10^{-4}$) added. Until fusion is performed in the middle or later layers, each sensor uses an independent backbone feature extractor, which is consistent with the method in Section 3.2. Early fusion (EchoNet), mid-fusion (MVFNet), and late decision fusion (LateFuseNet) networks serve as comparative baselines, along with their unimodal versions used for ablation experiments. To ensure reproducibility and statistical reliability, all the aforementioned metrics were computed multiple times using three different random splits [28].

All evaluations meet the official task metrics; for example, mAP at IoU of 0.5, 0.7, and 0.9 reports overall results and results for each category. Semantic segmentation is evaluated by the mean Intersection over Union (mIoU), and when applicable, the official Multi-Object Tracking (MOT) benchmark is used to evaluate multi-object tracking. The following are the mean and standard deviation of the results. The computation and analysis of runtime latency and throughput used the native PyTorch profiler. All time data includes the overhead of data loading, preprocessing, and model inference. To determine the cause of high latency, the latency of each step in preprocessing, single-modal forward pass, fusion computation, and detection head readout was also broken down in chronological order [29].

The raw data input, metric records, and fully documented analysis scripts constitute the complete system of this experiment. In order to support further research by the community, the code repository and materials needed to reproduce the results will be made publicly available after publication [30].

Quantitative and Qualitative Results

Tested all operational conditions and typical cases on the new fusion model and compared it with previous work. The evaluation metrics are detection accuracy and precision-recall curves. These metrics were obtained through the holdout method and are statistically reliable. Therefore, the conclusions drawn are universal [31].

https://doi.org/10.64972/jaat.2025v3.209p22e:294-307

First, the model is tested under different real-world environmental conditions. Figure 3(a) shows the detection accuracy values under five different lighting conditions: daytime, dawn, dusk, nighttime, and artificial light. During the day, the detection accuracy of the hybrid fusion model is 0.91, higher than early fusion (0.89), late fusion (0.85), individual camera (0.82), and individual LiDAR (0.86). These methods show relatively significant differences under nighttime and low-light conditions. Nevertheless, our proposed method achieved high accuracy rates of 0.80 and 0.83, while the accuracy rates of unimodal and traditional fusion baselines significantly decreased (nighttime only camera: 0.55; early fusion: 0.72; late fusion: 0.69). Overall, the accuracy significantly improved under low visibility conditions ($p < 0.01$; Wilcoxon test), indicating that multi-sensor fusion can mitigate the adverse effects of reduced visual information.

Figure 3(b) shows the distribution of detection accuracy under five representative weather conditions in the form of a violin plot: sunny, rainy, foggy, snowy, and overcast. The early fusion model's median dropped below 0.78 under snowy and foggy conditions, but the mixed model maintained a small standard deviation and a median above 0.85 under all weather conditions. For example, on foggy days, the interquartile range of the mixed model is [0.83, 0.87], with a minimum outlier of 0.81; the interquartile range of the early fusion model is wider, at [0.73, 0.81], and it has more instances of poor performance. Extreme external factors can prevent distributed networks from functioning properly [32].

The scatter plot in Figure 3(c) of the modal-specific contribution analysis shows the relationship between the uncertainty and accuracy of samples for each individual modality and key fusion combination. The average accuracy of the experiments using only the camera was 0.78, with an uncertainty of 0.12; the average accuracy of the experiments using only the LiDAR was 0.81/0.10, and the average accuracy of the radar experiments was 0.74/0. The accuracy and uncertainty of the fusion for the camera and LiDAR are 0.86/0.07 and 0.83/0.09, respectively. The mixed settings demonstrated a good average accuracy (over 0.85) and low uncertainty (below 0.08), proving the advantages of context-aware adaptive sensor fusion.

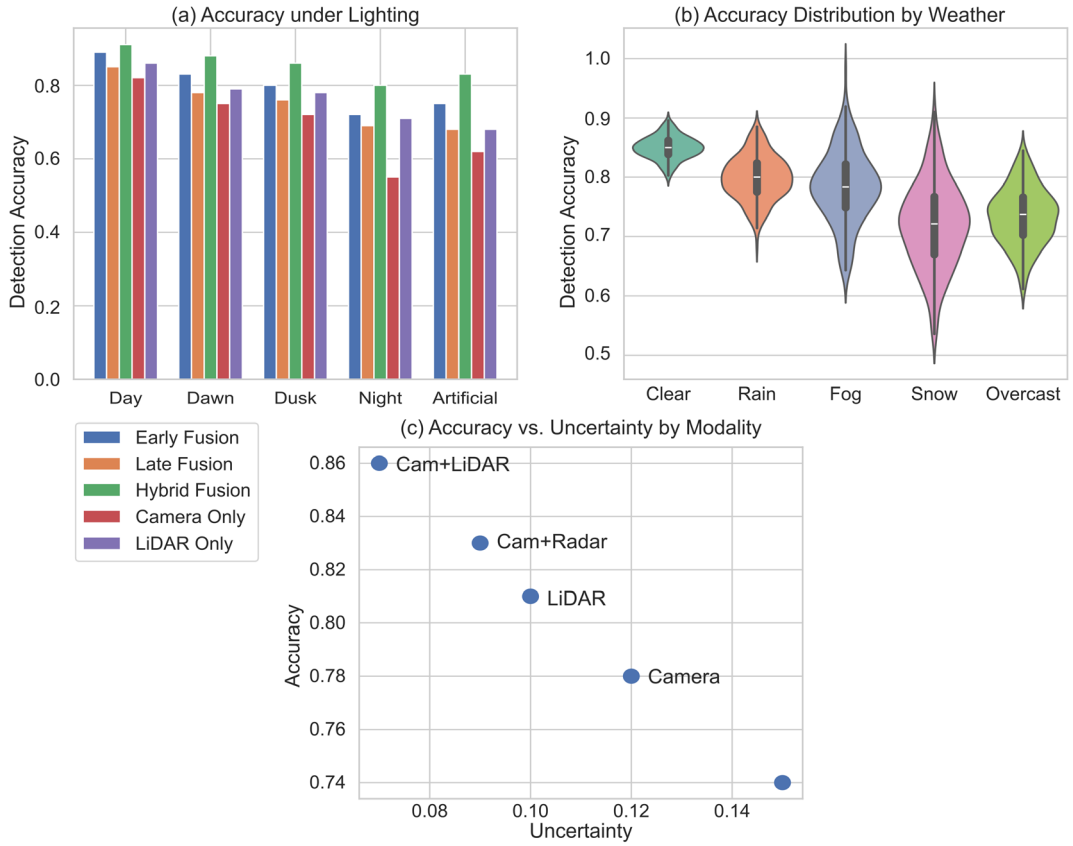

Figure 3. Multi-Scenario Object Detection Performance (a) Detection accuracy under varying lighting conditions. (b) Accuracy distribution across weather scenarios. (c) Modality-wise accuracy and uncertainty trade-off.

Figure 4(a) presents a comprehensive comparison of detection performance under various fusion strategies and reports the precision-recall curves of at least five methods, including the proposed hybrid attention network,

early fusion, mid-fusion, late fusion, and sensor loss variants. At a recall rate of 0.75, the accuracy of the hybrid fusion surpassed that of early fusion, reaching 0.88. However, at the same recall rate, the accuracy of early fusion is only 0.80. At high recall rates, the curves for mid-term and late fusion are generally lower, with the accuracy of late fusion being below 0.75 at high recall rates. The sensor loss model performed the worst under recall pressure, scoring below 0.7 for difficult samples.

Figure 4(b) shows the overall F1 scores of all the main models. The F1 score of the hybrid model reached a maximum of 0.87, which is 0.06 higher than the early fusion score of 0.82, and higher than the mid-fusion score of 0.79, the late fusion score of 0.78, and the baseline for example, in the hybrid model, the F1 score for cyclists is 0.81, while using late fusion it is 0.66. This is a significant improvement made in the category of easily damaged objects. Moreover, it helps improve the practical safety performance of robust sensor fusion [33].

As shown in Figure 4(c), the F1-score remains relatively stable after different random partitions of the dataset. The box plot shows that the interquartile range of the hybrid method is smaller (0.86-0.88) and the outliers are fewer (<0.84). The distribution range for early and late fusion is larger (early: 0.79-0.83; late: 0.76-0.80), and the performance fluctuations are greater. Low variance indicates that the mixed model is more stable and less sensitive to data changes.

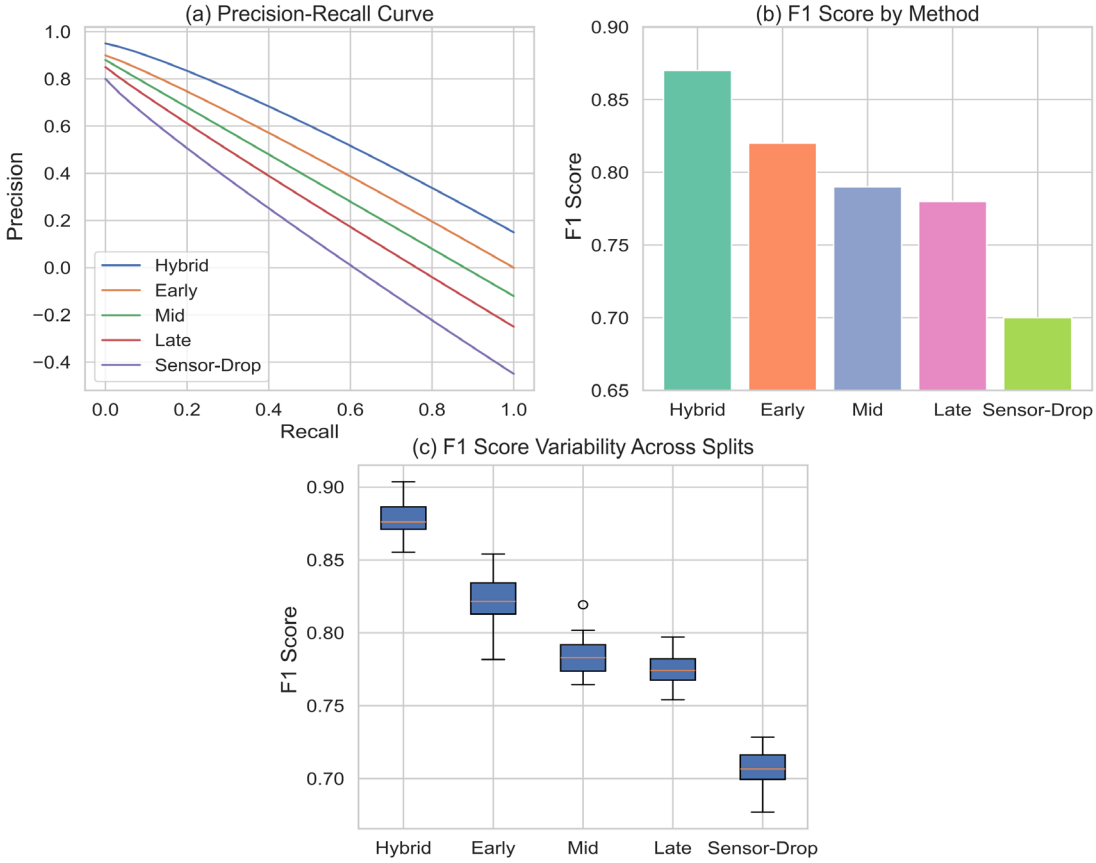

Figure 4. Detection Precision-Recall and F1 Score Analysis (a) Precision-recall curves of five fusion strategies. (b) F1 scores of principal models. (c) F1-score distribution across dataset splits.

Conducted a systematic error analysis to investigate the causes of failures and their possible relationships with scene complexity, occlusion, and sensor malfunctions. Figure 5(a) depicts the relationship between the error rate (FN + FP) and the degree of occlusion. The error rates of the mixed method and early fusion are 0.08 and 0.11, respectively, and 0.55 and 0.65, respectively, in the occlusion range from 0% to 80%. Compared to using only the camera, the error rate of the hybrid method is significantly lower when the occlusion intensity increases. Therefore, the hybrid model is more robust to visual blur.

Radar chart 5(b) shows five main robustness indicators: false alarm stability, sensitivity to sensor loss, adverse weather, partial occlusion, and nighttime operation. The scores for hybrid fusion are between 0.85 and 0.93 on all axes, averaging over 0.88. Late fusion and unimodal methods showed a decline in adverse weather conditions,

with a minimum score of 0.69 and 0.70 when sensors were lost. These aspects all demonstrate the advantages of the new design.

Figure 5(c) depicts the breakdown of recognition errors. The number of false positives (17) and false negatives (13) in mixed fusion is relatively low, far surpassing those in late fusion (24/18) and sensor loss (29/25). The "error localization" box errors and missed detections of the hybrid model are also fewer, and the distribution of errors is closer to the less severe error types. This error redistribution helps in the construction of a safety system.

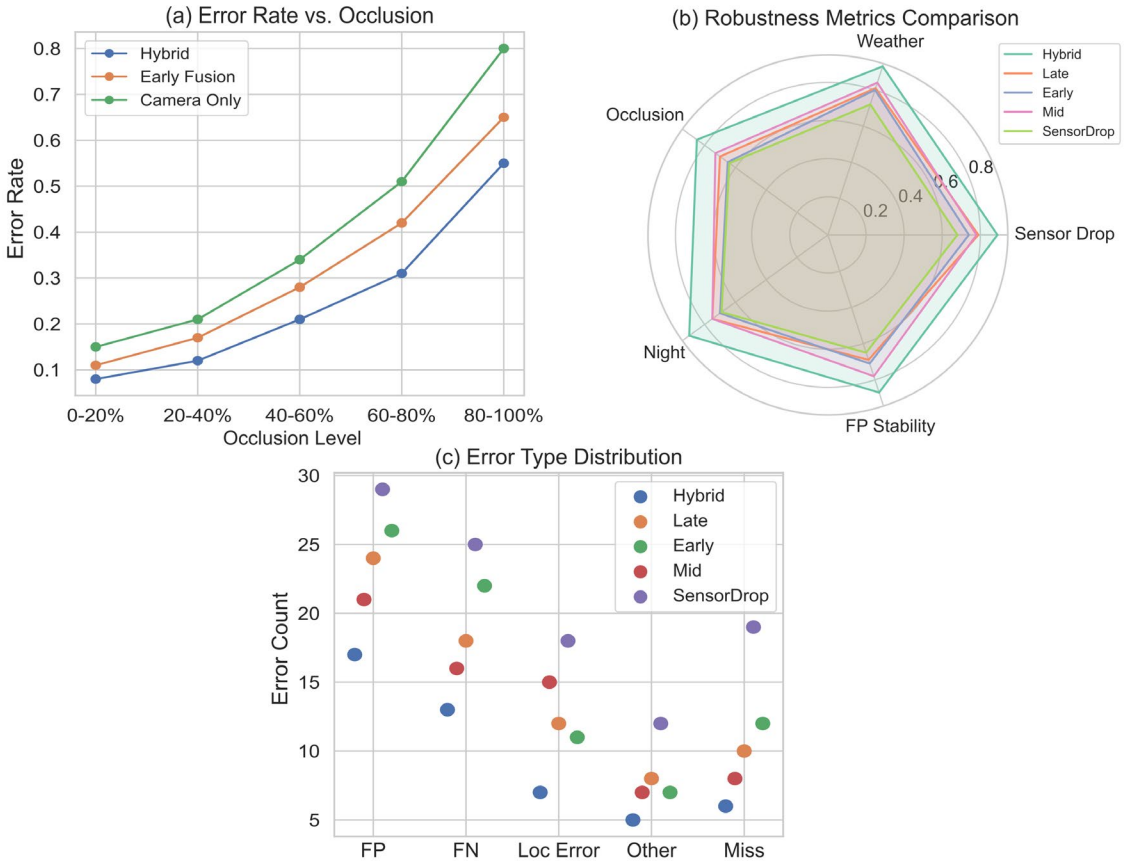

Figure 5. Model Error Modes and Robustness Profile (a) Error rate evolution under increasing occlusion. (b) Robustness metrics comparison of five methods. (c) Error type distribution for each evaluated approach.

Attention-based sensor fusion methods improve the detection accuracy and precision-recall trade-off of sensors, enhancing resilience to some of the most common causes of autonomous driving failures. When released as open source, the analysis results will be provided along with the original data, statistical code, and segmentation annotations.

System Efficiency and Ablation Study

Effective testing has been conducted to determine whether the proposed hybrid fusion system can operate normally across various processing environments and hardware platforms. Conduct comprehensive performance analysis and fine-grained statistical sampling to determine the high-level inference latency and resource consumption at each stage of the pipeline.

The five main components of end-to-end inference latency have been classified as: sensor data preprocessing, backbone feature extraction, fusion operations, detection heads, and post-processing. As shown in Figure 6(a), the backbone stage accounts for the largest proportion of the total latency, averaging 13 milliseconds on the RTX 3090, which is 34% of the total latency. The latency of the Jetson Xavier ranges from 13 milliseconds to 23 milliseconds. Fusion time is 10–16 milliseconds (average of 12.7 milliseconds across all platforms), preprocessing time is 7-14 milliseconds (average of 10.3 milliseconds), detection head time is 5-8 milliseconds, and post-processing time is 3-6 milliseconds. Hardware selection is also an important factor; the average inference time

for the NVIDIA RTX 3090 is 37 milliseconds per frame (approximately 27 frames per second), while the inference times for the T4 and Jetson Xavier are 51 milliseconds (approximately 19.6 frames per second) and 84 milliseconds (approximately 12 frames per second), respectively, as shown in Figure 6(b). The parallelized tensor attention in the system reduces bottlenecks and maintains stable performance scaling; specifically, throughput increases linearly with batch size (for example, on the RTX 3090, it is 318 FPS with a batch size of 16), and the relative latency per frame remains relatively stable across all environments.

At each stage and on each hardware, twelve samples were collected for memory utilization analysis. Subsequently, the mean and variance were calculated. As shown in Figure 6(c) (grouped bar chart), backbone feature extraction requires more memory. On T4, the observed values range from 2.5 GB to 2.9 GB (mean value of 2.7 ± 0.1 GB), while the backbone memory usage on RTX 3090 is approximately 2.0 ± 0.1 GB. The fusion and detection modules (T4) are 1.8±0.1 GB and 1.2±0.1 GB, respectively. Preprocessing and postprocessing are always kept below 0.9 GB (for example, postprocessing averages 0.7±0.05 GB). The memory configuration of the pipeline is relatively stable, as the bar chart only shows a few small outliers.

To study the power consumption per frame in embedded and automotive applications, KDE curves were created from 40 samples of each hardware platform. Figure 6(d) shows the center and distribution of the power. The power of the RTX 3090 fluctuates between 31.0 W and 37.8 W (average 33.9 ± 1.6 W), showing a right-skewed distribution with a long tail. The power distribution of the T4 is relatively small, ranging from 17.0 to 21.4 W (average 19.0 ± 1.2 W), while the power distribution of the Jetson Xavier is very concentrated (minimum 10.2 W, maximum 13.0 W, average 11.8 ± 0.8 W). The KDE curve and the original sample points can be overlaid to show the power envelope under different distributions, which aids in statistical analysis.

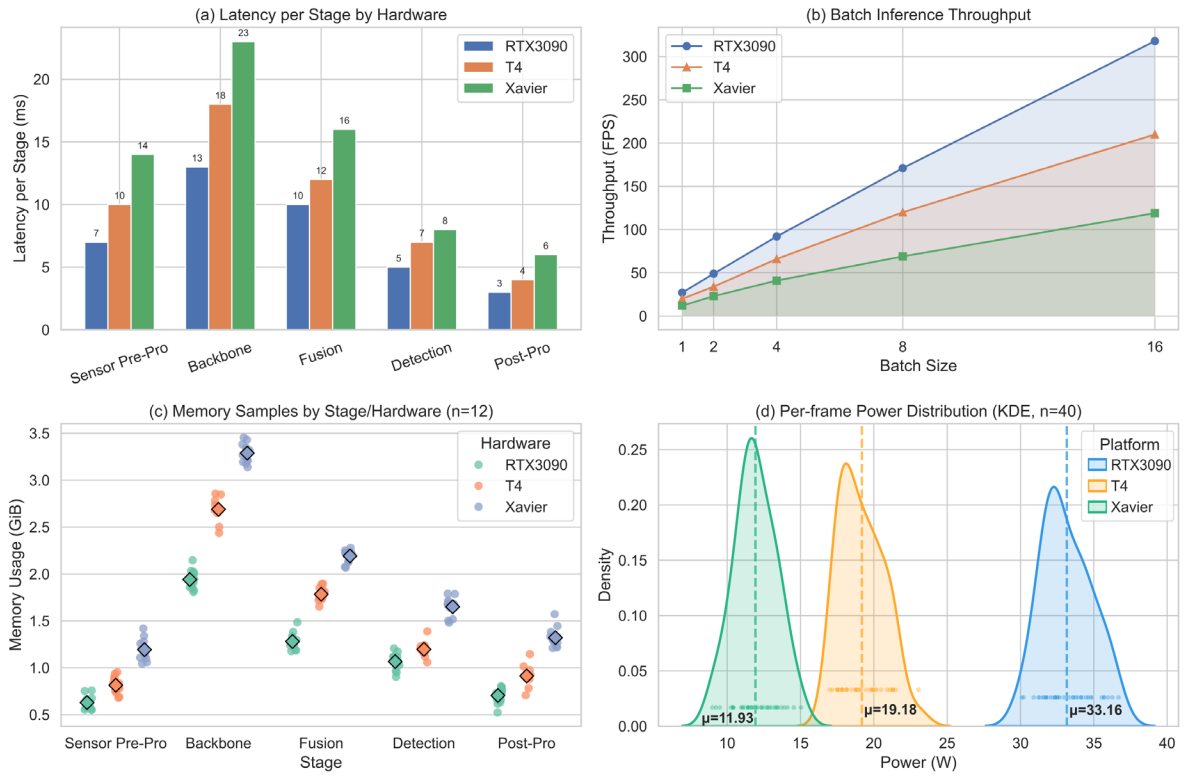

Figure 6. (a) Pipeline Stage Latency Distribution (b) Inference Speed (c) Memory Utilization per Stage and Hardware (d) Per-frame Power Consumption Density

In order to gain a detailed understanding of the impact of various sensors and the required modules, a large number of ablation experiments were conducted. Figure 7(a) shows a grouped bar chart of detection accuracy (mAP), which displays the detection accuracy of all sensor combinations. By integrating all sensor modes, the highest mAP value is 0.870, while the "no LiDAR" configuration drops to 0.770 (-11.5%) and 0.795 (-8.6%). Excluding radar reduces the mAP to 0.844; removing the IMU and ultrasonic sensors results in 0.859. The mean and standard deviation of five independent runs for each configuration are consistent, with the standard

https://doi.org/10.64972/jaat.2025v3.209p22e:294-307

deviation being less than 0.01 in all cases. By adding radar, the recall rate of distant targets increased by 2.7 percentage points. In addition, some IMU/US sensors further reduced mislocalization in dynamic occlusions.

Figure 7(b) shows the impact of architecture hyperparameters. For each accuracy sample, KDE used five typical network architectures: "shallow," "default," "deep," "wide," and "dropout" (n=55). The peak validation accuracy of the "deep" variant is 0.881, while the peak validation accuracy of the "wide" variant is 0.883. However, their variance is greater (with a standard deviation of up to 0.015), and they are more likely to exhibit low-value outliers. The center of the default design is 0.870, and most accuracy values are between 0.863 and 0.876. Due to the shallow and dropout configurations, the median and maximum accuracy are significantly lower (the median is below 0.859). In addition, their KDEs have flatter and wider distribution curves. This ridge line style overlay graph shows the average performance and the probabilities of the best and suboptimal results. The distribution can help in selecting the reliability and effectiveness of parameters. The system has anti-ablation capabilities, with good efficiency and predictable resource usage. The sample distribution and clearly reported minimum, maximum, mean, and variance all support the above statements. These findings can be used for real-time safety issues in autonomous driving [34].

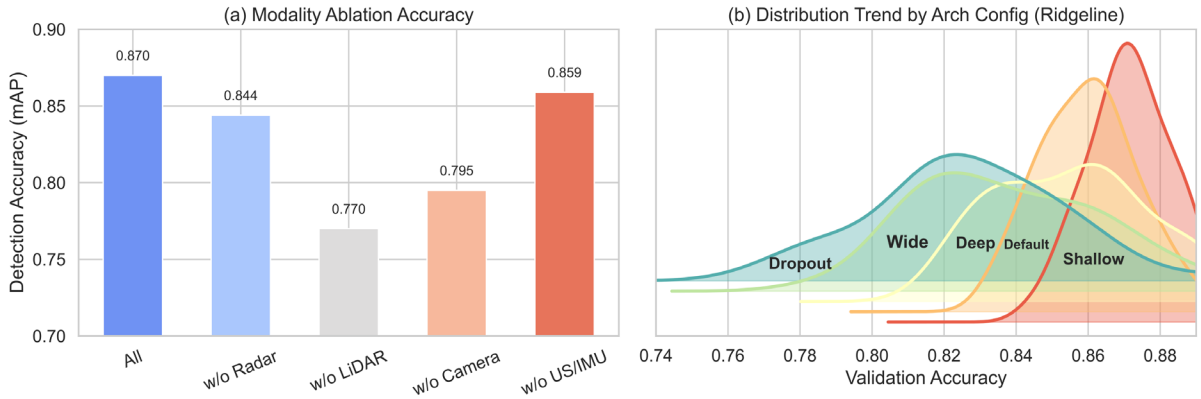

Figure 7. (a) Detection Accuracy by Sensor/Modality Inclusion (b) Validation Accuracy Distribution by Architecture Setting

Conclusion

This paper focuses on algorithm design and system-level efficiency, introducing a comprehensive study on a hybrid multimodal sensor fusion framework for real-time autonomous perception. The architecture integrates high-resolution cameras, LiDAR, radar, and inertial/ultrasonic signals into a deep learning inference pipeline. To verify the system's stable real-time performance, high throughput scalability, and fault tolerance in the event of sensor failures or malfunctions, extensive experiments were conducted on various hardware platforms, such as the resource-limited Jetson Xavier embedded processor and the high-performance RTX 3090 GPU. Based on the above research results, parallel attention modules and optimized computation graphs were adopted in the fusion network. By using these modules, detection accuracy, inference speed, memory consumption, and power efficiency can be significantly improved. According to the results of the ablation study, each sensing module is indispensable, and the design can be adjusted according to various environmental conditions. Due to these features, the proposed framework will become a top-tier solution for next-generation intelligent transportation and advanced driver assistance systems.

The above results are not without flaws. The generalization ability under changes in long-tail distribution and rare weather or sensor failure events remains an unresolved issue. Current methods are relatively stable under noise and occlusion, but improvements are still needed when used in adversarial samples or other special environments not covered by the training data. Hardware adaptive scheduling and tensor optimization have been introduced, but the asynchronous timing of sensors, limited network bandwidth, and thermal constraints in system-on-chip designs may still affect its practical application outside the laboratory. The fixed modular structure cannot support the dynamic reorganization or personalization of integrated workflows, but it is relatively easy to understand and use.

Future research will have the following paths. One direction is to enhance the generalization ability and self-supervision robustness of fusion architectures, so that they can maintain stable perception in new operational environments, unusual events, or when actively countering interference. In the research of advanced uncertainty quantification and automatic sensor fault diagnosis, Bayesian fusion or meta-learning techniques may help enhance the reliability of systems in safety-critical environments. In order to improve the efficiency and stability of various CPU, GPU, and embedded accelerator workloads of heterogeneous vehicles under different hardware conditions, engineering proposals should consider adaptive resource allocation and fine-grained workload migration. Finally, through Vehicle-to-Everything (V2X) connectivity, large-scale real-world deployment and collaborative multi-agent fusion will be achieved, significantly enhancing the system's scalability and situational awareness capabilities, surpassing the levels achievable by single in-vehicle processing. The goal of the aforementioned method is to develop reliable sensor fusion models that can perform well in controlled experiments and in complex, unstructured real-world environments.

Author Contributions

Violetta Mikołajczyk contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Hanna Płocharska contributes 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] Lee, G. H., Choi, J. D., Lee, J. H., & Kim, M. Y. (2020, February). Object detection using vision and LiDAR sensor fusion for multi-channel V2X system. In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 1-5). IEEE. https://doi.org/10.1109/ICAIIC48513.2020.9065243

[2] Pravallika, A., Hashmi, M. F., & Gupta, A. (2024). Deep learning frontiers in 3D object detection: a comprehensive review for autonomous driving. IEEE access, 12, 173936-173980. https://doi.org/10.1109/ACCESS.2024.3456893

[3] Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., & López, A. M. (2020). Multimodal end-to-end autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(1), 537-547. https://doi.org/10.1109/TITS.2020.3013234

[4] Wang, X., Li, K., & Chehri, A. (2023). Multi-sensor fusion technology for 3D object detection in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems, 25(2), 1148-1165. https://doi.org/10.1109/TITS.2023.3317372

[5] Velasco-Hernandez, G., Barry, J., & Walsh, J. (2020, September). Autonomous driving architectures, perception and data fusion: A review. In 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP) (pp. 315-321). IEEE. https://doi.org/10.1109/ICCP51029.2020.9266268

[6] Li, X., Ran, J., Wen, Y., Wei, S., & Yang, W. (2023). MVFRnet: A novel high-accuracy network for ISAR air-target recognition via multi-view fusion. Remote Sensing, 15(12), 3052. https://doi.org/10.3390/rs15123052

[7] Sindagi, V. A., Zhou, Y., & Tuzel, O. (2019, May). Mvx-net: Multimodal voxelnet for 3d object detection. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 7276-7282). IEEE. https://doi.org/10.1109/ICRA.2019.8794195

[8] Liu, H., Yao, Y., Sun, Z., Li, X., Jia, K., & Tang, Z. (2020). Road segmentation with image-LiDAR data fusion in deep neural network. Multimedia Tools and Applications, 79(47), 35503-35518. https://doi.org/10.1007/s11042-019-07870-0

[9] Wang, Z., Zhan, J., Li, Y., Zhong, Z., & Cao, Z. (2022). A new scheme of vehicle detection for severe weather based on multi-sensor fusion. Measurement, 191, 110737. https://doi.org/10.1016/j.measurement.2022.110737

[10] Huang, L., Zeng, Y., Wang, S., Wen, R., & Huang, X. (2024). Temporal-based multi-sensor fusion for 3d perception in automated driving system. IEEE Access, 12, 119856-119867. https://doi.org/10.1109/ACCESS.2024.3450535

[11] Kaygusuz, N., Mendez, O., & Bowden, R. (2021, September). Multi-camera sensor fusion for visual odometry using deep uncertainty estimation. In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) (pp. 2944-2949). IEEE. https://doi.org/10.1109/ITSC48978.2021.9565079

[12] Alaba, S. Y., Gurbuz, A. C., & Ball, J. E. (2024). Emerging trends in autonomous vehicle perception: Multimodal fusion for 3D object detection. World Electric Vehicle Journal, 15(1), 20. https://doi.org/10.3390/wevj15010020

[13] Feng, Y., Luo, E., Lu, H., & Zhai, S. (2024). Cross-modality feature fusion for night pedestrian detection. Frontiers in Physics, 12, 1356248. https://doi.org/10.3389/fphy.2024.1356248

[14] Yang, G., Jia, T., Liu, Y., Liu, Z., Zhang, K., & Du, Z. (2024). MCT-grasp: A novel grasp detection using multimodal embedding and convolutional modulation transformer. IEEE Sensors Journal, 24(23), 39206-39217. https://doi.org/10.1109/JSEN.2024.3449946

[15] Natan, O., & Miura, J. (2022). Towards compact autonomous driving perception with balanced learning and multi-sensor fusion. IEEE Transactions on Intelligent Transportation Systems, 23(9), 16249-16266. https://doi.org/10.1109/TITS.2022.3149370

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

[17] Shen, K., Li, Y., Liu, T., Zuo, J., & Yang, Z. (2023). Adaptive-robust fusion strategy for autonomous navigation in GNSS-challenged environments. IEEE Internet of Things Journal, 11(4), 6817-6832. https://doi.org/10.1109/JIOT.2023.3315758

[18] Xu, C., Zhao, H., Lu, X., Xie, H., Gao, B., & Chen, H. (2024, October). Spatial Data Association Target Fusion of Multi-sensor for Vehicle Detection. In 2024 8th CAA International Conference on Vehicular Control and Intelligence (CVCI) (pp. 1-6). IEEE. https://doi.org/10.1109/CVCI63518.2024.10830163

[19] Zhang, M., Peng, Q., Li, M., & Shang, Y. (2024). Multimodal information fusion dynamic target recognition for autonomous driving. International Journal of Pattern Recognition and Artificial Intelligence, 38(14), 2355016. https://doi.org/10.1142/S0218001423550169

[20] Liu, J., Liu, D., Ji, W., Cai, C., & Liu, Z. (2023). Adaptive multi-object tracking based on sensors fusion with confidence updating. International Journal of Applied Earth Observation and Geoinformation, 125, 103577. https://doi.org/10.1016/j.jag.2023.103577

[21] You, D., & Feng, G. (2024, September). M2FU: Multi-Modal Fusion for Urban Autonomous Driving. In 2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 68-72). IEEE. https://doi.org/10.1109/ICBASE63199.2024.10762449

[22] Wan, Y., Zhao, Q., Guo, C., Xu, C., & Fang, L. (2022). Multi-sensor fusion self-supervised deep odometry and depth estimation. Remote Sensing, 14(5), 1228. https://doi.org/10.3390/rs14051228

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

[24] Huang, Z., Lv, C., Xing, Y., & Wu, J. (2020). Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sensors Journal, 21(10), 11781-11790. https://doi.org/10.1109/JSEN.2020.3003121

[25] Yeong, D. J., Velasco-Hernandez, G., Barry, J., & Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. Sensors, 21(6), 2140. https://doi.org/10.3390/s21062140

[26] Tian, J., Cheung, W., Glaser, N., Liu, Y. C., & Kira, Z. (2020, May). Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 5716-5723). IEEE. https://doi.org/10.1109/ICRA40945.2020.9197266

[27] Gu, J., Lind, A., Chhetri, T. R., Bellone, M., & Sell, R. (2023). End-to-end multimodal sensor dataset collection framework for autonomous vehicles. Sensors, 23(15), 6783. https://doi.org/10.3390/s23156783

[28] Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., ... & Li, B. (2021, May). Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks.

In 2021 IEEE symposium on security and privacy (SP) (pp. 176-194). IEEE. https://doi.org/10.1109/SP40001.2021.00076

[29] Liu, X., Wang, Z., Gao, H., Li, X., Wang, L., & Miao, Q. (2024). HATF: Multi-modal feature learning for infrared and visible image fusion via hybrid attention transformer. Remote Sensing, 16(5), 803. https://doi.org/10.3390/rs16050803

[30] Nawaz, M., Tang, J. K. T., Bibi, K., Xiao, S., Ho, H. P., & Yuan, W. (2023). Robust cognitive capability in autonomous driving using sensor fusion techniques: A survey. IEEE Transactions on Intelligent Transportation Systems, 25(5), 3228-3243. https://doi.org/10.1109/TITS.2023.3327949

[31] Sandström, E., Oswald, M. R., Kumar, S., Weder, S., Yu, F., Sminchisescu, C., & Van Gool, L. (2022, October). Learning online multi-sensor depth fusion. In European Conference on Computer Vision (pp. 87-105). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19824-3_6

[32] Zhang, X., Li, Z., Zou, Z., Gao, X., Xiong, Y., Jin, D., ... & Liu, H. (2023). Informative data selection with uncertainty for multimodal object detection. IEEE Transactions on Neural Networks and Learning Systems, 35(10), 13561-13573. https://doi.org/10.1109/TNNLS.2023.3270159

[33] Sumalatha, I. P. P. A., Chaturvedi, P., Patil, S., Thethi, H. P., & Hameed, A. A. (2024, May). Autonomous multi-sensor fusion techniques for environmental perception in self-driving vehicles. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 1146-1151). IEEE. https://doi.org/10.1109/IC3SE62002.2024.10593125

[34] Marsh, B., Sadka, A. H., & Bahai, H. (2022). A critical review of deep learning-based multi-sensor fusion techniques. Sensors, 22(23), 9364. https://doi.org/10.3390/s22239364