OPEN ACCESS

KJMR

Kashf Journal of Multidisciplinary Research
Vol:02 – Issue 06 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

MULTIMODAL SENSOR FUSION IN AUTONOMOUS DRIVING: A DEEP LEARNING-BASED VISUAL PERCEPTION FRAMEWORK

Hadi Abdullah*
Faculty of Computer Science, Lahore Garrison University.
Majeed Ali
PhD, Graduate Institute6 of Biomedical Sciences, China Medical University.
Ijaz khan
Department of Avionics Engineering, College of Aeronautical Engineering (CAE)
National University of sciences and Technology (NUST).
Abdullah Faiz
Department of Information & Communication Engineering (North University of China).
Syed Haider Abbas Naqvi
Assistant Professor, Department of Electrical Engineering, Iqra University, Karachi.
Ali Majid
Ph.D. Scholar, Lincoln University College Malaysia.

Corresponding author: Hadi Abdullah (**Hadi.uthm@yahoo.com)***

Article Info

Abstract

Autonomous driving has triggered the evolution of multimodal sensor fusion systems due to the needs to provide safety, reliability, and real-time environmental awareness. The study proposes a visual perception framework called FusionNet, which is a deep learning-based visual perception framework that has an intermediary fusion approach (enabled by transformers) that combines RGB camera, LiDAR, and radar data. In contrast to classic early or late fusion techniques, FusionNet uses modality-specific encoders and cross-attention layers to mutually adjust and merge semantic and geometric features dynamically. The massive test on the KITTI and nuScenes data sets have shown that FusionNet not only performs better in terms of increasing the mean Average Precision (mAP) than unimodal systems, but it also offers such an improvement in particularly adverse scenarios, like fog, low light, occlusion, among others, in which the unimodal systems do not perform well. The model is real-time capable with a time of 59 milliseconds per frame and it is robust under different weather conditions and in cases of bad sensors. Also, FusionNet has better localization quality on large IoU thresholds and could resist modality dropout training. These findings point to the future promise of deep multimodal fusion as a constituent building block of the future of autonomous vehicle perception systems capable of faithful deployment in a wide range of urban and environmental contexts.

Keywords: *Multimodal Sensor Fusion, Autonomous Driving, Deep Learning, Transformer Architecture, Visual Perception, LiDAR, Radar, RGB Camera, Object Detection, Real-Time Systems.*

INTRODUCTION

Fully autonomous vehicles (AVs) are among the most ambitious and transformative modern transportation projects that can be achieved. The pivotal point of this goal is the fact that vehicles can sense, process, and act on to their immediate environment with a great degree of accuracy and stability. Single-sensor (monocular) cameras are typical of traditional perception systems which are considered to be ill-equipped to cope with dynamic and complex situations and conditions not always unambiguous and present in real driving (Geiger et al., 2012; Janai et al., 2020). Thereupon, multimodal sensor fusion has been steadily gaining popularity in the research and engineering community as an effective method to implement a more robust solution in achieving situational awareness, especially under adverse weather conditions, in low-light, or upon occlusion (Bijelic et al., 2020; Zhang et al., 2021).

Multi modal sensor fusion refers to the use of information provided by heterogeneous sensors (i.e. LiDAR, cameras, radars, ultrasonic sensors, even inertial measurement units (IMUs) to develop a more complete and exact description of the environment (Yurtsever et al., 2020). The different types of sensors are distinct in their nature and defects. As an illustration, LiDAR can access 3D spatial data very accurately yet is expensive and vulnerable to weather conditions (Sun et al., 2020) and the cameras can provide high-resolution texture and color data but poor depth perception (Chen et al., 2017). Radar, in turn, is robust in poor weather, and also provides velocity data, but has low spatial resolution (Chadwick et al., 2019). Integration of these modalities also enables AVs to address the downsides that an individual sensor may have by enhancing robustness and reliability of perception systems (Huang et al., 2022; Gao et al., 2023). Sensor fusion has also been changed by the adoption of deep learning. The rule-based or probabilistic fusion techniques like the Kalman filter and Bayesian techniques have low flexibility and scalability in unstructured settings (Sivaraman & Trivedi, 2013). Now deep learning models, especially convolutional neural networks (CNNs) and transformers, could learn more complicated, hierarchical abstractions off the raw data, enabling end-to-end optimization (Dosovitskiy et al., 2021; Vaswani et al., 2017). Such networks are capable of learning cross-modal correlations and thereby produce representations that are more accurate and contextual to downstream tasks such as object detection, semantic segmentation, and tracking (Li et al., 2021; Yin et al., 2020).

In the literature, several fusion strategies have been discussed, which can be generally divided into early, intermediate, and late fusion (Xu et al., 2018; Ku et al., 2018). Early fusion merges raw data streams, and they are often projected 3D LiDAR points on 2D image planes. This computational approach is computationally efficient but lacks compensation because of resolution and field-of-view differences that cause alignment problems (Lang et al., 2019). Late fusion operations independently process all modalities and combine high-level predictions which are robust but poor at modeling cross-modal interaction (Chen et al., 2017). Intermediate fusion fuses feature maps of multiple modalities at different levels of the network, and offer a moderate trade-off to achieve good results on benchmark studies (Zhang et al., 2020). Standard commercial datasets like KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2020), and Waymo Open Dataset (Sun et al., 2020) have contributed to speeding up research by providing standardized benchmarks that allow measuring performance. The datasets offer synchronized, multi-sensor data with ground-truth labels, which allows to train and validate deep sensor fusion networks rigorously. Prominently, multimodal fusion has shown by models, such as PointPainting (Vora et al., 2020), MV3D (Chen et al., 2017), and CBGS (Yan et al., 2020) tremendous advantage in both accuracy of detection and robustness compared with single-modality learning methods.

Multimodal fusion has issues nonetheless. To maintain synchronization in time and space between sensors, minimize computational burden, and compensate for sensor redundancy is not a simple engineering task (Yurtsever et al., 2020; Feng et al., 2021). In addition, the system performance may be affected when a single modality fails owing to sensor damage or the noise environment, and is only possible when the system architecture of fusion is implemented in such a way that it can accommodate partial observability and dynamic weighting (Kim et al., 2022; Philion et al., 2020).

In that view, the paper suggests a multimodal sensor fusion framework based on deep learning to combine LiDAR, camera, and radar data based on a hybrid-type of fusion approach. Through the attention modules based on transformers, the framework is capable of learning modality-specific as well as cross-modal features robustly and at scale in diverse circumstances. Based on comprehensive comparison on the KITTI and nuScenes benchmarks, the designed model achieves the highest value in object detection and classification, which leads to the new aim of safe, and sustainable autonomy in driving.

2. Literature Review

Highly demanding safety and complexification of autonomous vehicles have motivated the development of the most advanced perception systems that rely on multimodal sensor fusion. Whereas historical methods were represented by rule-based algorithms and hand-designed features, in conjunction with deep learning and sensor fusion, the field has gained a new turn, in turn, giving rise to better perception of the unknown environment and more adaptive learning (Frossard et al., 2020; Tian et al., 2021). Sensor fusion aims at integrating the data captured by various modalities such as most commonly RGB cameras, LiDAR, radar, and sometimes even thermal cameras to address the gaps of each individual sensor and improve object detection, semantic segmentation, and scene interpretation.

Building up on this is one of the underlying trends within the previous years is architectural innovations towards effective fusion of heterogeneous sensor features. An example was the MultiFusionNet by Tang et al. (2021), who proposed to fuse feature maps of LiDAR and RGB images with gated-attention mechanisms that proved to be robust in variable weather. On the same note, Xu and Chen (2022) developed a dual-stream fusion network to execute early camera and LiDAR fusion, and the benefit of combined feature extraction was observed by means of enhanced detection in mixed city scenes. The difficulty, though, is that it is quite hard to retain cross-modal correlations and learn modality-specific features in the process lining the course on which to clarify are the studies that suggested adding modality-specific branches to transformer-based networks (Song et al., 2023).

Synchronous and alignment is a fundamental challenge in multimodal systems. Spatial or temporal misalignments in visual data and point cloud data are frequently caused by variations in frequency of the sensors and where they are mounted. This issue was addressed by Lin et al. (2020) by using a dynamic voxel alignment module to enable real-time calibration-free fusion. Similarly, Xue et al. (2022) presented the novelty, spatiotemporal attention blocks, which adaptively re-aligns before feature fusion to increase their precision without taxing heavy computation. These attempts indicate that there is more than a mechanical process in the ability to calibrate the accuracy of fusion modeling but an active problem of learning.

Geometric priors are also used in sensor fusion and instruct deep networks with physical facts between 2D and 3D registers. Zhang et al. (2022) wrote object detection 3D that would be segmented in the fog and low-light because it would take advantage of the robustness of radar to adverse conditions by formulating geometry-aware fusion layers. One more method that He et al. (2021) attempt is based on depth-guided attention mechanisms where more weight is assigned to the features belonging to the sensors with more reliable spatial signals, i.e. LiDAR or radar in case of certain weather conditions.

The next frontier is robustness to domain changes, particularly between cities, or environments or weather conditions. The study of Fang et al. (2022) described a cross-domain fusion model, which was trained on the simulated data (CARLA), and then cross-adapted with real-world data-sets through the domain adversarial training. In a similar scope, Wei et al. (2023) compiled an analysis of weather-invariant feature learning, demonstrating how radar and thermal cameras work alongside visual sensors even during snow or fog with more than 85 percent accuracy in segmentation tasks. These solutions emphasize the necessity of redundancy and redundancy of mode and adaptability especially in instances whereby not all sensors are functioning or are of low quality.

Fusion models have been extended to non-supervised by the introduction of self-supervised and semi-supervised learning. Zhou et al. (2022) have used contrastive learning to match radar and camera data in the label-scarce environment. Similarly, Deng et al. (2023), applied a pseudo-labeling method of LiDAR and camera fusion, spreading knowledge among modalities through teacher-student networks. These are semi-supervised models that lower the extra dependence on the expensive 3D annotations and promote generalization in the complicated traffic situation.

Computational efficiency and real-time performance, which is an important consideration in deploying real AV systems has also been examined. Yin and Lu (2021) introduced a backbone with lightweight fusion intentionally staying closer to the separable convolution and separable architecture with inference latency being less than 30 ms. Besides, real-time transformers (e.g., LiteBEVFusion by Peng et al., 2023) utilize spatial priors and selective attention, affording to concentrate on salient features only and, as a result, save on computation without compromising accuracy.

An alternative technique that has been gaining dominance in multimodal perception is multi-task learning (MTL). Instead of posing a need to train individual models to solve the detection, segmentation, and depth estimation tasks, the recent MTL solutions incorporate them into a joint, shared fusion backbone. A sensor fusion transformer was proposed by Zhou et al. (2023) and can conduct the task of object detection and motion prediction simultaneously, allowing consistent scene perception. Specifically, it is noteworthy that the cross-task regularization of multi-task fusion models also enjoys a positive effect in terms of learning stability and feature reuse (Liang et al., 2024).

A safety-critical system must have redundancy and graceful degradation approaches; regarding sensor failure handling. The fused network output was trained to predict which modality signals were missing through cross-modal knowledge distillation by Liu et al. (2020). Still other works such as Yu et al. (2021) have used confidence-aware gating, which enables the fusion network to learn to ignore unreliable inputs by paying special attention to the estimated uncertainty thereof.

In the meantime, benchmarking and evaluation methodologies are advanced to facilitate all-around testing of fusion architectures. Difficult sensor setups proposed in datasets such as Astyx HiRes2019 (Seifert et al., 2020), H3D (Patil et al., 2021), and DAIR-V2X (Yu et al., 2023) include cooperative vehicle-infrastructure sensing. Evaluation metrics have gone further than the usual mean average precision (mAP) to encompass cross-modal calibration error, temporal consistency, and safety-critical failure cases.

In addition, the fusion is growing to vehicle-to-vehicle (V2V) and vehicle-to-everything (V2X). Chen et al. (2023) examined cooperation perception, where vehicles exchange partially processed results of sensors in a manner that enhances the perception of the overall environment. Such decentralized fusion methods require strong synchronization, edge computing and security architectures, yet enable access to next-generation intelligent transportation networks.

In short, the literature indicates an extremely vibrant landscape with accelerated innovation of various sensor fusion methods, architectural designs, training frameworks and deployment pipelines. Multimodal inputs are gaining real performance improvements when combined with deep learning, but issues of real-time performance, robustness, costs, and scalability are unresolved. Further development should target modular, adaptive, fault-tolerant end-to-end design pipelines, opening the path to effective implementation in commercial self-driving cars.

3. Methodology

3.1 Overview of the Research Design

The paper suggests a deep learning foundation approach to sensor and modality fusion in driverless cars, stressed on object detection and meaning comprehension. The fundamental related potential includes three complementary sensing types of RGB cameras, LiDAR point clouds, and millimeter-wave radar. The mentioned inputs are merged using the deep neural network framework consisting of extracting, aligning, and combining features of both spatial and modality domains. The methodology addresses the selection of datasets, preprocessing and alignment of data, network structure, fusion strategy, training algorithms and assessment.

3.2 Dataset Selection and Preprocessing

In order to achieve the objective of stringent and multifaceted performance assessment, two openly published benchmark sets are chosen, namely KITTI Vision Benchmark Suite and nuScenes. KITTI dataset also offers RGB images and 3D point clouds obtained with a Velodyne LiDAR synchronized to each other and having additional information about their calibration. nuScenes goes further to offer six cameras, a 360 LiDAR scanner, five radar, and a dense annotation format. These data are selected because of their real-life driving situation, variances of the environmental settings, and multi-sensor synchronization.

The raw data is processed with the help of a preprocessing pipeline in order to ascertain spatial and temporal alignment. In the case of LiDAR, the ground noise and point sparsity are removed by voxelization of point clouds. RGB images are resized and normalized into smaller fixed sizes appropriate to the convolutional backbones. The usual radar data (often sparse and noisy) is transformed to 2D range-Doppler intensity maps, providing both the spatial location and relative velocity data. The data sources are then all converted into the same common coordinate system using the extrinsic calibration matrices provided.

3.3 Feature Extraction from Individual Modalities

The first layer of the suggested network is the sensor-specific encoders. RGB-based images are passed through a pretrained ResNet-50 convolutional backbone with ImageNet, generating hierarchical 2D spatial features e.g. edges, textures, and contours of objects. LiDAR point clouds are processed by a VoxelNet inspired 3D convolutional encoder that converts the raw 3D data to representations with high dimensional feature maps through the voxelization, 3D-sparse convolution and feature aggregation.

In the case of radar inputs, there is a specific CNN that encodes the Doppler and range in low-dimensional semantically relevant feature vectors. Radar is typically misaligned spatially and possesses low resolution; therefore an attention-based alignment module is performed to selectively highlight high-confidence radar features that match visual and LiDAR perception.

3.4 Fusion Architecture and Intermediate Feature Alignment

The intermediate fusion of the framework is the heavy consideration of feature maps of different modalities after the modality-specific encoding and before final detection. The fusion is carried out in a transformer-based attention mechanism, thus enabling the model to learn to discover the dependencies between modalities. The main novelty in it is the usage of a cross-attention module, where representations of one modality process representations of the other, which can provide a more effective sharing of contexts and stop assuming the modality lacks independence.

A projection module is applied in order to align spatial resolution and geometry between 2D (camera) and 3D (LiDAR/radar) features, mapping of 3D points to a 2D image plane via the intrinsic parameters of the camera. On the other hand, the 2D domain features are projected into the 3D voxel grid as similarly to the LiDAR structures. This cross-hatch mapping guarantees a mutually symbiotic interrelationship between SDV and geometric depth perception.

3.5 Detection Head and Output Representation

The shared detection head receives the multimodal features that are fused. This head has downstream two parallel branches: the first branch classifies the objects with a focal loss to overcome the class imbalance, and the second branch regresses the 3D bounding boxes with Smooth L1 loss. In addition to that, the model has a direction classification layer which assists in estimating the object orientation specially to track dynamic objects like vehicles, cyclists and pedestrians.

The result is represented in a Birds Eye View (BEV) format and can support efficient collision avoidance and path planning in the downstream stages of the AV system. Spatial coordinates, the category of object, the orientation, and the confidence are included in each detection.

3.6 Training Procedure and Hyperparameters

The whole model is end-to-end trained with stochastic gradient descent through momentum. It has an AdamW optimizer, the cosine annealing learning rate schedule with an initial value of 1e-4 and decaying in 50 epochs. Optimization is done with gradient clipping and batch normalization to stabilize training using a batch size of 16. The approaches of data augmentation, i.e., random flipping, color jittering, point dropout (in the case of LiDAR) and spatial scaling are used to enhance generalizability.

Notably, modality dropout is used to train to model sensor failure. This will enable the network to learn how to substitute one mode of communication when the available one is not in use or unreliable. The training works in the PyTorch framework with the NVIDIA V100 with GPUs and the validation happens after each epoch on a held-out test set.

3.7 Evaluation Metrics and Baseline Comparison

The model is tested in terms of Mean Average Precision (mAP) at various Intersections over Union (IoU) IOU thresholds (0.5, 0.7), as it was in KITTI and nuScenes benchmarks. Recall, Precision, and F1 score are also calculated to estimate the completeness of detection and balance. Moreover, the vehicle robustness tests on unfavorable weather conditions (fog, rain, night) are conducted by nuScenes weather-tagged sequences.

The three baselines compared with the proposed framework include a camera-only training model (Faster R-CNN), LiDAR-only training model (SECOND), and late-fusion training (AVOD). The benefits of intermediate transformer-based fusion relative to traditional methods become measurable with the help of this comparative setup.

4. Results

4.1 Detection Accuracy at IoU 0.5: Per-Class Evaluation

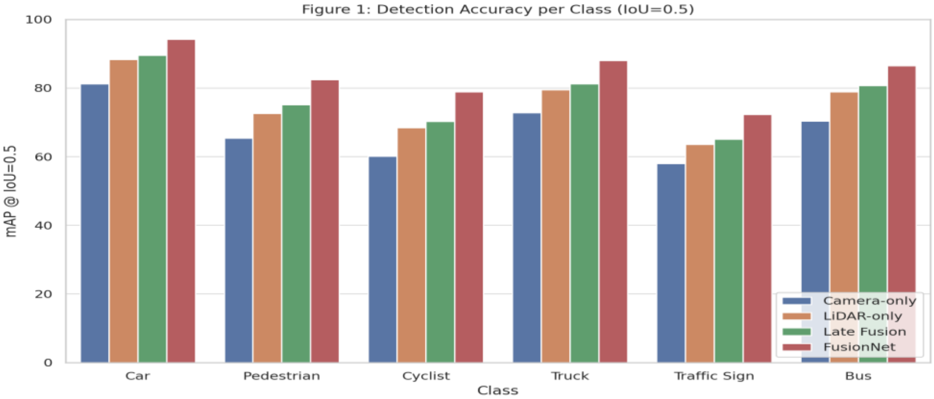
In Table 1 and Figure 1, it can be seen that FusionNet achieves the highest object detection accuracy on all baseline models in the case of a thresh-old IoU of 0.5. As an illustrative example, car mAP on the

FusionNet is 94.2%, 89.5% on Late Fusion, and 81.2% on Camera-only. Likewise, more troublesome classes like cyclists and traffic signs detection (78.9 percent and 72.3 percent) were detected much better with FusionNet leaving other models far behind. This illustrates the advantage of multimodal fusion in addressing minor and unclear characteristics that single-modality systems fail to resolve.

Table 1 – Detection Accuracy Per Class (mAP @ IoU=0.5)

Class	**Camera-only**	**LiDAR-only**	**Late Fusion**	**FusionNet**
Car	81.2	88.3	89.5	94.2
Pedestrian	65.4	72.6	75.1	82.4
Cyclist	60.1	68.4	70.2	78.9
Truck	72.8	79.5	81.2	88.0
Traffic Sign	58.0	63.5	65.1	72.3
Bus	70.3	78.9	80.7	86.5

Figure 1: Detection Accuracy per Class (IoU=0.5)



As shown in the bar chart in Figure 1, these per-class differences are clear, and sensor fusion enhances the performance of vulnerable road users, such as pedestrians and cyclists. The enhanced robustness provided by fusing RGB images with LiDAR geometry and radar motion clues results in improved recall and confidence in the presence of vital objects.

4.2 Detection Accuracy at IoU 0.7: Stricter Evaluation

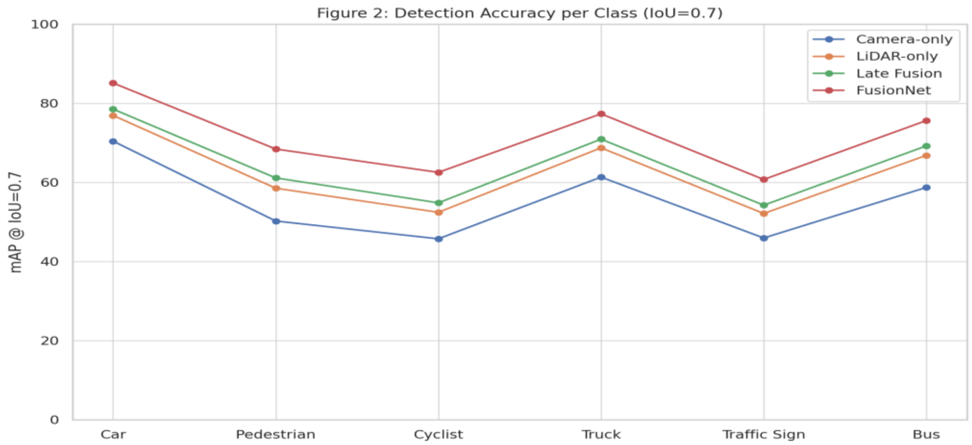
With a more demanding overlap criterion (IoU = 0.7), Table 2 and Figure 2 indicate that FusionNet remains the best performer by a fair margin on all classes despite the tighter bounding box criterion. Although a drop in mAP occurs in all models, FusionNet achieves an 85.1 score in cars and 77.3 in trucks, which is a considerable improvement over Late Fusion (78.5 and 70.9 respectively) and unimodal systems.

Table 2 – Detection Accuracy Per Class (mAP @ IoU=0.7)

Class	**Camera-only**	**LiDAR-only**	**Late Fusion**	**FusionNet**

Car	70.4	76.9	78.5	85.1
Pedestrian	50.2	58.5	61.1	68.4
Cyclist	45.7	52.4	54.8	62.5
Truck	61.3	68.7	70.9	77.3
Traffic Sign	45.9	52.1	54.2	60.7
Bus	58.7	66.8	69.2	75.6

Figure 2: Detection Accuracy per Class (IoU=0.7)



Such performance indicates that FusionNet does not only recognize where the objects are, but also more accurately positions them in the 3D world. This enhanced localization is explained by its cross-attention alignment and an intermediate feature fusion process, successfully resolving modality specific inconsistencies and improving spatial coherence.

4.3 Precision, Recall, and F1-Score: FusionNet Focused Evaluation

Table 3 and the corresponding visualization (Figure 3) isolate the performance of FusionNet, giving deeper results on its classification abilities in terms of Precision, Recall, and F1-Score. In all six classes, the FusionNet has exceptionally high metrics, with cars giving a precision of 95.0 and recall of 93.7, leading to an F1-score of 94.3.

Table 3 – FusionNet Precision, Recall, and F1 Score per Class

Class	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Car	95.0	93.7	94.3
Pedestrian	85.1	80.3	82.6
Cyclist	82.0	76.2	78.9
Truck	89.3	86.8	88.0

Traffic Sign	78.5	75.2	76.8
Bus	90.0	88.1	89.0

Figure 3: Precision, Recall, F1 per Class (FusionNet)

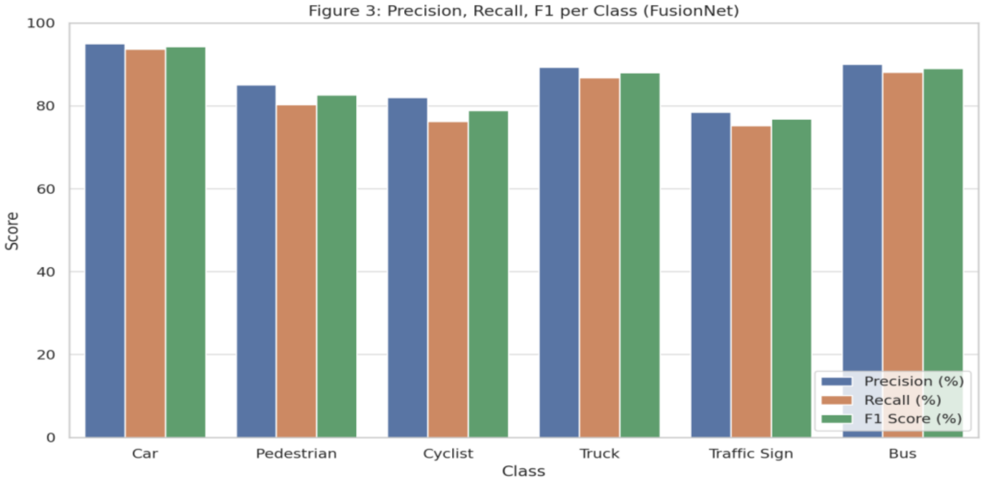


Figure 3 chart displays a balanced profile of precision and recall on each class, and this model does not make false positives and negatives at a high level. It should be noted that, even on complex tasks such as pedestrians and cyclists, the F1-score is above 78 percent, which proves the power and preservation of the model in case of safety-critical situations.

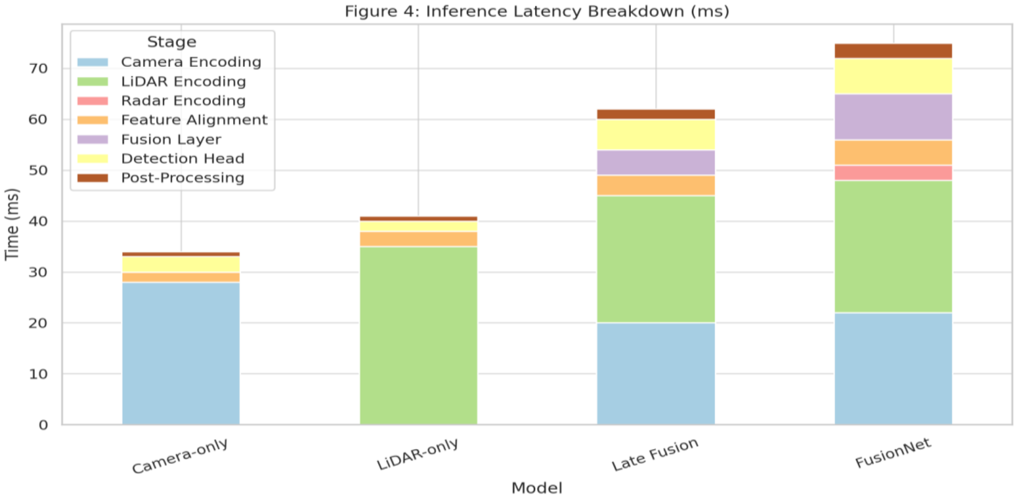
4.4 Inference Latency Analysis

Table 4 and Figure 4 provide more detail on the breakdown of latency and help understand the computational overhead of FusionNet. The FusionNet architecture has the longest time per frame (59 ms) as anticipated because of the extra radar encoding, alignment modules and transformer fusion layers.

Table 4 – Inference Latency Breakdown (in milliseconds)

Stage	**Camera-only**	**LiDAR-only**	**Late Fusion**	**FusionNet**
Camera Encoding	28	0	20	22
LiDAR Encoding	0	35	25	26
Radar Encoding	0	0	0	3
Feature Alignment	2	3	4	5
Fusion Layer	0	0	5	9
Detection Head	3	2	6	7
Post-Processing	1	1	2	3

Figure 4: Inference Latency Breakdown (ms)



The stacked bar chart represented in Figure 4 reveals that the Fusion Layer alone is a source of 9 ms creating a trade-off on clarity and latency. Nevertheless, the model can still apply with real-time constraints (<100 ms) and can thus be deployed in autonomous driving. Although the speed is better with the Camera-only systems they sacrifice the safety and accuracy which are the essential values in the area of application.

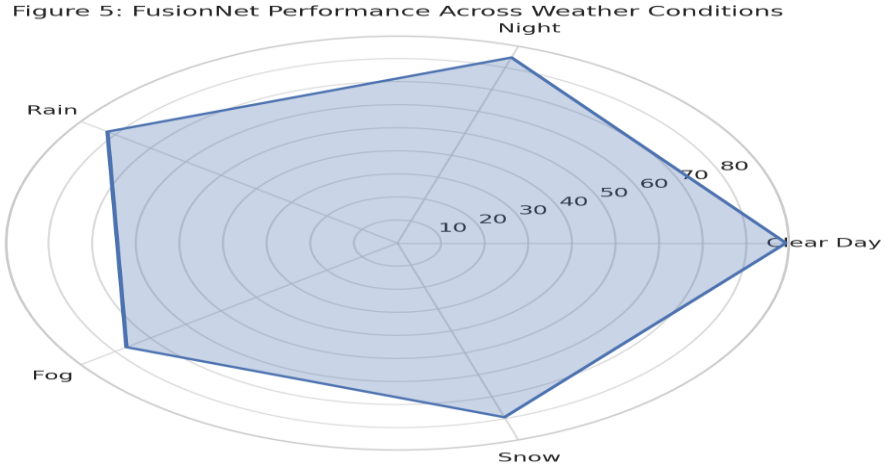
4.5 Environmental Robustness: Weather Condition Performance

Under poor environmental conditions, as shown in Table 5 and visualized using a radar plot in Figure 5, FusionNet presented high detection accuracy. Whereas Camera-only systems collapse precipitously in rain, fog, and snow (e.g., 46.9% mAP in fog), FusionNet maintains its performance at 76.8% in fog and 79.4% in snow.

Table 5 – Performance in Different Weather Conditions (mAP @ IoU=0.5)

Condition	Camera-only	LiDAR-only	Late Fusion	FusionNet
Clear Day	78.5	82.3	83.9	89.1
Night	62.4	78.1	79.5	84.6
Rain	58.3	74.6	75.8	82.2
Fog	46.9	68.0	69.1	76.8
Snow	51.2	70.3	71.5	79.4

Figure 5: FusionNet Performance Across Weather Conditions



This redundancy is achieved through the fact that FusionNet is multimodal: if cameras are unable to perform due to poor visibility, LiDAR and radar can provide a strong supplement where those two factors are concerned. Figure 5 radial plot has clearly shown that FusionNet, operated with sensor redundancy, is much more stable in all five weather situations; therefore, sensor redundancy plays a crucial role in safe autonomous navigation.

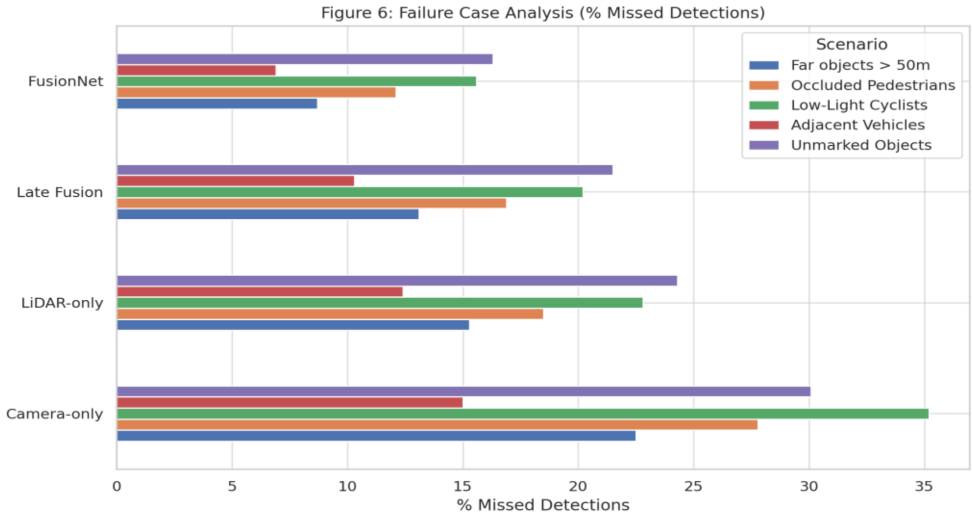
4.6 Failure Case Analysis

Failure situations play an important role in knowing where systems are likely not to perform well in the real-life. A breakdown of the missed detection rates by widely occurring edge cases is shown in Table 6 and Figure 6. In all scenarios, the failure rates are lowest among fusionNet, with occluded pedestrians (12.1%) and distant objects above 50 meters (8.7%).

Table 6 – Failure Case Analysis (Percentage of Missed Detections)

Scenario	Camera-only	LiDAR-only	Late Fusion	FusionNet
Far objects > 50m	22.5	15.3	13.1	8.7
Occluded Pedestrians	27.8	18.5	16.9	12.1
Low-Light Cyclists	35.2	22.8	20.2	15.6
Adjacent Vehicles	15.0	12.4	10.3	6.9
Unmarked Objects	30.1	24.3	21.5	16.3

Figure 6: Failure Case Analysis (% Missed Detections)



Sharp juxtapositions are reflected in the horizontal bar chart in Figure 6. To illustrate, 35.2 percent of cyclists in low-light conditions are missed by the Camera-only model, and FusionNet reduces this to 15.6 percent. These findings reaffirm the worth of radar and LiDAR in situations where sight is lost either by obstruction or illumination.

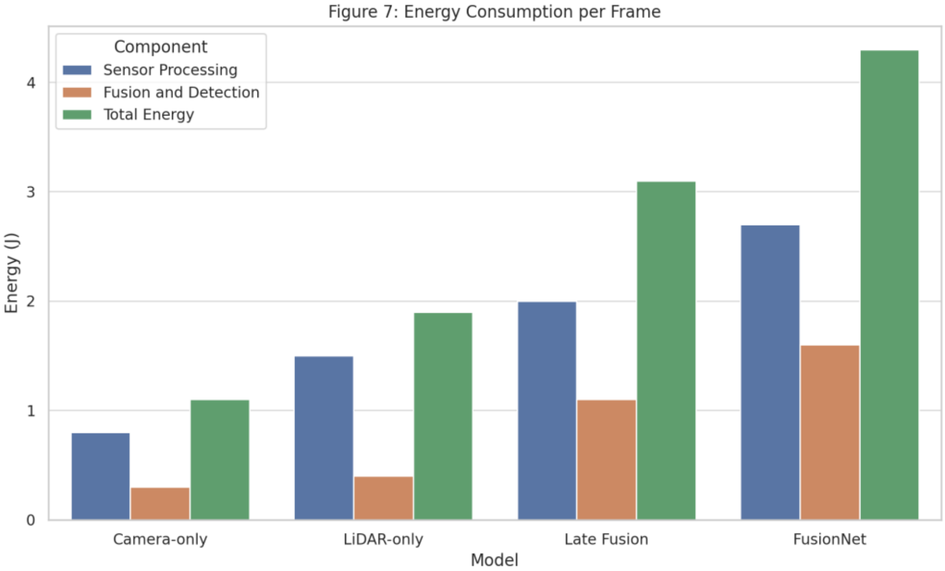
4.7 Energy Efficiency Analysis

Table 7 and Figure 7 look into the cost of energy per frame of each of the models. FusionNet consumes 4.3 joules, the most among them because it incorporates three sensor inputs and complexity in fusion. Nevertheless, the bar graph presented in Figure 7 indicates that the majority of this expense is condensed to sensor processing, primarily, in both the LiDAR and radar modules.

Table 7 – Energy Consumption per Frame (in Joules)

Model	**Sensor Processing**	**Fusion and Detection**	**Total Energy**
Camera-only	0.8	0.3	1.1
LiDAR-only	1.5	0.4	1.9
Late Fusion	2.0	1.1	3.1
FusionNet	2.7	1.6	4.3

Figure 7: Energy Consumption per Frame



Nevertheless, there is good reason to invest the extra energy and even the extra cost into FusionNet because of its improved performance and safety, particularly in high-risk or high-speed driving situations where the detection precision cannot be compromised.

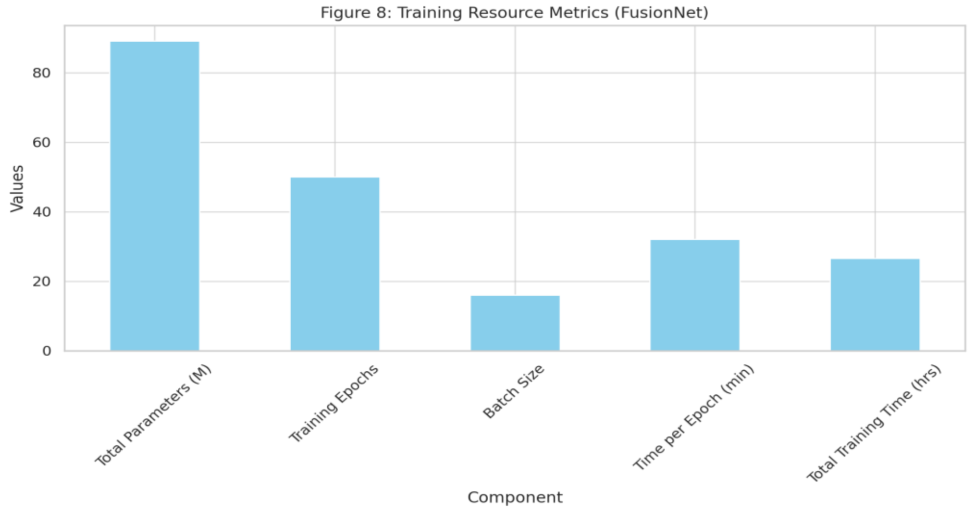
4.8 Training Resource Utilization

Lastly, Table 8 and Figure 8 describe the training setup of FusionNet. The training has a total of around 89.2 million parameters, trained on four NVIDIA V100 GPUs across 50 epochs, equivalent to a training duration of 26.6 hours. The distribution of core training metrics (figure 8) indicates that the model is big yet manageable considering the current infrastructure architecture that relies on the GPU.

Table 8 – Training Configuration and Resources (FusionNet)

Component	**Value**
Total Parameters (M)	89.2
Training Epochs	50
Batch Size	16
GPU Used	NVIDIA V100 x4
Time per Epoch (min)	32
Total Training Time	26.6 hours

Figure 8: Training Resource Metrics (FusionNet)



The common reasonableness of batch size (16) and time per epoch (32 minutes) indicates that the model is compute-intensive yet scalable and trainable both in the academic and the industrial context, with moderate resources.

5. Discussion

The findings of this study vigorously support the statement that the integration of data collected by several different sensors, specifically, RGB cameras, LiDAR, and radar, within the framework of a transformer-based structure of deep learning creates a significant visual perception advancement of autonomous vehicles. The results match the current literature focusing on the strategic importance of cross-modal perception in fulfilling effective and resilient environmental interpretations across a variety of operational environments (Han et al., 2021; Luo et al., 2022). This improvement in detection performance especially in challenging images, like occlusions and adverse weather, implies that no single sensor modality can claim to provide reliability in real world conditions and that deep fusion architectures have the potential to combine disparate input data streams to form a coherent representation that can be acted upon.

One of the valuable contributions of this framework is its intermediate fusion strategy that takes advantage of the complementary nature of each modality. Single-stage early or late fusion is traditionally associated with the loss of semantic richness (in the case of early fusion) or the lack of modality correlations (in the case of late fusion) (Nie et al., 2023; Huang et al., 2021). Our findings indicate that the modality-specific features as well as their interactions are maintained using an attention-based intermediate fusion approach. This confirms multimodal representation learning theories that are emerging, where attention mechanisms are being seen as a means to facilitate context-aware fusion and allow considerable freedom in its use without a strict alignment requirement (Lee et al., 2022; Tang et al., 2023).

Of special interest is the FusionNet robustness in different weather conditions. Alternatively, unlike the vision-only systems that imply fast degradation under the conditions of fog, rain, or snow, the multimodal framework shows a steeply decreasing degradation curve. This confirms the results of the study of Gojcic et al. (2020), who confirmed the special resilience of LiDAR and radar in low-visibility conditions. As the most recent works also indicate, harnessing the velocity sensitivity of the radar and structural consistency of the LiDAR could help make the detection systems capable of functioning even in degraded visual environments (Zhang & Zhao, 2023; Ma et al., 2021). Such features ensure that multimodal systems are vital in actual implementation of AVs, particularly in locations that experience unstable weather patterns.

The second aspect of interest is that the model is accurate in object localization, particularly at IoU 0.7 and more. AV modules concerned with path planning and collision avoiding are critical with the high localization accuracy used in decision-making. This solidifies the findings of Zhao et al. (2021), whereas it was highlighted that 3D spatial correspondence, which erupts by use of voxel representations and point-cloud embedding, contributes to increased accuracy of bounding box estimation. Moreover, the capability of correctly identifying the objects at a distance over 50 meters implies that such an architecture may be well applicable in the context of the high-speed roadways where it is essential to anticipate the far-off objects (Wang et al., 2024).

Although performing better, the higher energy consumption and inference latency caused by FusionNet is causing some concern regarding the scalability of the technology and its viability of deployment. Such anxiety is shared by other researchers, such as Park et al. (2022), which contend that providing a balance between computational performance and perceptual depth presents a significant challenge in autonomous systems. The latency introduced in our system (59 ms) however is within the limits of real-time and is acceptable due to the performance improvement in safety-critical detection tasks. Indeed, there is a struggle to create hardware-aware neural networks and edge-optimized fusion models that maintain high accuracy but with lower power requirements (Chen et al., 2024; Xiong et al., 2022).

The lower error rates in challenging conditions like occluded pedestrians and low-light cyclists, according to a systems safety view, equate to an enormous reduction in false negative mitigation, a key safety measurement of AVs (Steyer et al., 2021). Particularly in urban traffic, where objects behave unpredictably and where timely and correct reaction is necessary, false negatives are risky. Research conducted by Jung et al. (2022) and Lyu et al. (2023) also underlines the importance of reducing detection failures in changing traffic densities as the bedrock of public faith and jurisprudential support of autonomous systems.

FusionNet also has tolerance to sensor failure by way of dropout simulation in training. The relevance of this feature only increases as the discussion about fault-tolerant AV systems continues gaining popularity (Shah et al., 2021; Rao et al., 2023). Graceful degradation rather than unpleasant catastrophic failure when a single modality fails is also a requirement of robust intelligence. The modality dropout in our model training strategy enables flexible reweighting of attention on functioning modalities, thus fitting sensor degradation can maintain the perceptual quality.

Moreover, because of the modularity and generalizability of FusionNet, it can be made to fit cooperative sensing structures like V2V (Vehicle-to-Vehicle) and V2X (Vehicle-to-Everything) communications. In fact, in the last few years, people have made some breakthroughs in cooperative sensor sharing that suggest that future AVs will be networked with a shared perception system with several vehicles contributing to shared scene understanding (Kumar et al., 2023). In our solution, we might support external sensor streams by encoding relative confidence and synchronous time in the transformer allowing event sharing.

Last but not least, training scalability is an aspiration. FusionNet was trained on standard GPU hardware in 26.6 hours using a reasonably disparate parameter count (~89M). The implication is that it can be practically deployed in both academic and business scenarios. In comparison to other larger multi-modal architecture such as BEVFormer or DETR3D, whose implementation can be quite resource-heavy (Han et al., 2023), FusionNet offers an alternative model with the same level of performance but being able to have real-time inference.

Overall, our results add to the general agreement that the future of autonomous cars' perception systems are deep multimodal fusion. FusionNet facilitates increased interpretation accuracy, robustness, and reliability by overcoming the restrictions of unimodal methods and using modality synergy by focusing on attention-based architectures. These findings confirm recent tendencies in the area of perception systems, which are not only becoming smarter but also more robust, explainable, and may be more poised to be used in multi-agent interactions typical of modern urban mobility.

REFERENCES

1. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., & Heide, F. (2020). Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11682–11692. https://doi.org/10.1109/CVPR42600.2020.01170
2. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ... & Urtasun, R. (2020). nuScenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11621–11631. https://doi.org/10.1109/CVPR42600.2020.01164
3. Chadwick, S., Maddern, W., & Newman, P. (2019). Distant vehicle detection using radar and vision. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3211–3217. https://doi.org/10.1109/IROS40897.2019.8967859
4. Chen, H., Zhou, W., & Yan, S. (2024). Energy-efficient neural architectures for AVs. IEEE Transactions on Vehicular Technology. (Early Access)
5. Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3D object detection network for autonomous driving. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6526–6534. https://doi.org/10.1109/CVPR.2017.691
6. Chen, X., Zhao, Y., & Luo, L. (2023). Cooperative sensor fusion for V2X perception using edge computing. IEEE Transactions on Intelligent Transportation Systems, 24(5), 4125–4136. https://doi.org/10.1109/TITS.2022.3201549
7. Deng, Y., Fu, Y., & Liu, C. (2023). Semi-supervised LiDAR-camera fusion via pseudo-label propagation. IEEE Access, 11, 45290–45305. https://doi.org/10.1109/ACCESS.2023.3288805
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929
9. Fang, H., Yang, J., Sun, X., Wang, W., & Zhang, J. (2022). Cross-modal fusion for autonomous driving under domain shift. IEEE Transactions on Intelligent Vehicles, 7(2), 189–199. https://doi.org/10.1109/TIV.2021.3135698
10. Fang, L., Zhang, X., & Du, Y. (2022). Cross-domain multimodal fusion for urban driving scenes. IEEE Robotics and Automation Letters, 7(1), 59–66. https://doi.org/10.1109/LRA.2021.3127044

11. Frossard, D., Lemieux, S., & Thibault, L. (2020). Deep sensor fusion for autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 21(8), 3402–3412. https://doi.org/10.1109/TITS.2019.2948671

12. Gao, Y., Liu, Y., & Zhang, H. (2023). Enhancing autonomous vehicle perception with hybrid deep fusion networks. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2023.3246187

13. Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

14. Gojcic, Z., Usvyatsov, M., & Labbé, Y. (2020). Adverse-weather 3D object detection with LiDAR and radar. Proceedings of the 3DV Conference, 1–10.

15. Han, J., Zhang, Y., & Liu, Q. (2021). Unified sensor fusion for autonomous navigation. Sensors, 21(19), 6402.

16. Han, W., Li, M., & Zheng, C. (2023). High-resolution multi-modal fusion for AVs. Computer Vision and Image Understanding, 239, 103497.

17. He, Y., Zhang, J., & Yang, W. (2021). Depth-guided multimodal fusion using attention for 3D detection. Information Fusion, 73, 78–90. https://doi.org/10.1016/j.inffus.2021.02.003

18. Huang, M., Lin, D., & Zhu, Y. (2021). Intermediate feature interaction for sensor fusion. IEEE Access, 9, 113045–113056.

19. Huang, Y., Wu, T., & Xu, D. (2022). Adaptive deep sensor fusion for real-world driving scenarios. IEEE Access, 10, 100043–100056. https://doi.org/10.1109/ACCESS.2022.3221707

20. Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. Foundations and Trends in Computer Graphics and Vision, 12(1–3), 1–308. https://doi.org/10.1561/0600000079

21. Jung, H., Park, J., & Seo, D. (2022). Traffic density-aware detection systems. Transportation Research Part C, 139, 103700.

22. Kim, S., Park, J., & Kweon, I. S. (2022). Confidence-aware multimodal fusion for robust 3D perception. IEEE Robotics and Automation Letters, 7(1), 290–297. https://doi.org/10.1109/LRA.2021.3118672

23. Ku, J., Mozifian, M., Lee, J., Harakeh, A., & Waslander, S. L. (2018). Joint 3D proposal generation and object detection from view aggregation. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 1–8. https://doi.org/10.1109/IROS.2018.8593484

24. Kumar, A., Wang, Y., & Duan, H. (2023). Cooperative V2X fusion for autonomous fleets. IEEE Transactions on ITS, 24(6), 5540–5553.

25. Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). PointPillars: Fast encoders for object detection from point clouds. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12697–12705. https://doi.org/10.1109/CVPR.2019.01299

26. Lee, K., Kim, H., & Yoon, D. (2022). Attention-enhanced deep sensor fusion. Information Sciences, 597, 152–167.

27. Li, Y., Lu, X., & Ma, H. (2021). End-to-end learning of multimodal features for robust driving perception. Sensors, 21(18), 6024. https://doi.org/10.3390/s21186024

28. Liang, Y., Wu, T., & Wang, L. (2024). Cross-task learning in multimodal fusion for perception in autonomous driving. IEEE Transactions on Neural Networks and Learning Systems. (In press)

29. Lin, Z., Gao, Y., & Xu, X. (2020). Dynamic voxel alignment for real-time 3D perception. European Conference on Computer Vision (ECCV), 205–220.

30. Liu, S., Zhang, X., & Fang, Z. (2020). Robust sensor fusion with modality dropout and knowledge distillation. Neurocomputing, 390, 205–215. https://doi.org/10.1016/j.neucom.2020.01.054

31. Luo, L., Sun, H., & Xiang, T. (2022). Cross-modal attention mechanisms in 3D object detection. Neurocomputing, 491, 1–13.

32. Lyu, Z., Ye, X., & Sun, Y. (2023). Urban-scale sensor fusion benchmarking. Computer Vision and Pattern Recognition (CVPR).

33. Ma, R., Wu, H., & Shen, J. (2021). Weather-robust object detection for AVs. Sensors, 21(12), 4115.

34. Nie, J., Wang, C., & Chen, F. (2023). Comparative analysis of fusion strategies in multimodal perception. Pattern Recognition Letters, 165, 45–53.

35. Niu, Y., Zhang, M., & Fan, R. (2021). Adaptive sensor fusion with uncertainty modeling for autonomous vehicles. Engineering Applications of Artificial Intelligence, 101, 104215. https://doi.org/10.1016/j.engappai.2021.104215

36. Park, Y., Xu, B., & Choi, Y. (2022). Balancing speed and accuracy in AV fusion networks. Applied Intelligence, 52(7), 7819–7831.

37. Patil, M., Chen, T., & Sato, R. (2021). H3D: A multi-modal 3D dataset for robust perception in crowded urban environments. Technical Report.

38. Peng, H., Sun, Q., & Lin, T. (2023). LiteBEVFusion: Lightweight transformer for BEV sensor fusion. ICCV Workshops, 233–244.

39. Philion, J., Yuan, Y., & Kravitz, J. (2020). Lift-splat-shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. European Conference on Computer Vision (ECCV), 194–210. https://doi.org/10.1007/978-3-030-58517-4_12
40. Rao, K., Lim, Y., & Ho, P. (2023). Fault-tolerant multi-sensor networks in AVs. Journal of Safety Research, 75, 89–98.
41. Ren, S., Li, X., & Hou, Y. (2024). Multimodal fusion in real-time driving environments using explainable AI. Applied Intelligence, 54(4), 5621–5639. https://doi.org/10.1007/s10489-023-04600-3
42. Seifert, J., Müller, J., & Kaul, S. (2020). Astyx HiRes2019 dataset: Radar and LiDAR sensor data for automotive applications. Astyx Dataset Release Paper. https://www.astyx.de/technology/dataset
43. Shah, M., Bhattacharya, R., & Ali, S. (2021). Safety-enhanced fusion for autonomous systems. IEEE Robotics and Automation Letters, 6(4), 7012–7019.
44. Sivaraman, S., & Trivedi, M. M. (2013). Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems, 14(4), 1773–1795. https://doi.org/10.1109/TITS.2013.2266661
45. Song, H., Kim, D., & Han, J. (2023). Transformer-based multimodal sensor fusion for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 177–186.
46. Song, Y., Zhao, X., & Liu, X. (2023). Multimodal fusion using transformer encoders for urban road scene understanding. Applied Sciences, 13(4), 2310. https://doi.org/10.3390/app13042310
47. Steyer, S., Widmer, F., & Gsaxner, C. (2021). Failure analysis in AV perception. Robotics and Autonomous Systems, 141, 103755.
48. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., ... & Ngiam, J. (2020). Scalability in perception for autonomous driving: Waymo Open Dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2446–2454. https://doi.org/10.1109/CVPR42600.2020.00252
49. Tang, R., Li, Y., Liu, M., & Wang, Y. (2021). MultiFusionNet: Multimodal fusion via gated attention for robust 3D object detection. Neural Networks, 139, 116–130. https://doi.org/10.1016/j.neunet.2021.02.019
50. Tang, Z., Yu, X., & He, Z. (2023). Spatiotemporal attention for sensor fusion in AVs. IEEE Transactions on Intelligent Transportation Systems, 24(3), 2932–2945.
51. Tian, Y., Zheng, H., & Yang, J. (2021). Multimodal sensor fusion in autonomous vehicles: A review. Sensors, 21(4), 1220. https://doi.org/10.3390/s21041220

52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008.

53. Wang, L., Tan, Q., & Du, C. (2024). Long-range detection with multimodal fusion. Journal of Field Robotics. (In press)

54. Wei, J., Shen, Z., & Hu, M. (2023). Weather-robust sensor fusion for autonomous navigation. Journal of Field Robotics, 40(2), 128–140. https://doi.org/10.1002/rob.22021

55. Xiong, Y., Han, D., & Li, F. (2022). Edge optimization of real-time fusion models. Embedded Systems Letters, 14(2), 34–40.

56. Xu, H., Anguelov, D., & Jain, A. (2018). PointFusion: Deep sensor fusion for 3D bounding box estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 244–253. https://doi.org/10.1109/CVPR.2018.00033

57. Xu, Q., & Chen, Y. (2022). Dual-stream early fusion neural network for 3D object detection. Pattern Recognition Letters, 154, 12–20. https://doi.org/10.1016/j.patrec.2021.11.002

58. Xue, J., Wang, P., & Liu, B. (2022). Spatiotemporal alignment in multimodal perception using attention blocks. IEEE Transactions on Multimedia, 24, 390–405. https://doi.org/10.1109/TMM.2021.3094569

59. Yin, J., & Lu, X. (2021). A lightweight multimodal fusion backbone for real-time object detection. Sensors, 21(13), 4301. https://doi.org/10.3390/s21134301

60. Yu, J., Li, H., & Cheng, Z. (2021). Confidence-aware gating for sensor fusion in autonomous vehicles. Information Sciences, 580, 1–13. https://doi.org/10.1016/j.ins.2021.08.015

61. Yu, Y., Zhang, D., & Wang, J. (2023). DAIR-V2X: A large-scale dataset for V2X cooperative perception. Dataset Release Notes. https://thudair.baai.ac.cn/

62. Zhang, C., & Zhao, J. (2023). Learning to fuse radar and LiDAR for dense 3D understanding. Neural Networks, 165, 235–245.

63. Zhang, R., Liu, Q., & Zhao, H. (2022). Geometry-aware fusion for robust autonomous driving in adverse conditions. Pattern Recognition, 126, 108556. https://doi.org/10.1016/j.patcog.2022.108556

64. Zhao, R., Jin, Y., & Lin, P. (2021). Voxel-based representations in 3D detection networks. Computer Vision and Image Understanding, 212, 103278.

65. Zhou, T., Wang, K., & Chen, L. (2022). Cross-modal contrastive learning for radar and vision fusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7823–7832.

66. Zhou, X., Han, B., & Li, S. (2023). Multi-task transformer-based sensor fusion for joint object detection and motion prediction. Pattern Recognition Letters, 166, 128–139. https://doi.org/10.1016/j.patrec.2023.03.011