Journal of Applied Automation Technologies

Research Paper

Multi-Sensor Fusion-Based Lane Detection Using Convolutional Neural Networks

Szymon Sadlak[1, *], Walerian Wąsik[1] and Cyryl Borowski[2]

[1] Faculty of Mechatronics and Automation, Cracow University of Technology, Krakow, 31-155, Poland
[2] Faculty of Mechatronics and Automatic Control, University of Applied Sciences in Nysa, Nysa, 48-300, Poland
*Corresponding author: szymon.sa@pk.edu.pl

Abstract. Precise and stable lane recognition is necessary for the current smart transportation system to operate safely and for vehicle location and path stability. Conventional vision-based techniques struggle in real-world situations like poor illumination, bad weather, or partial occlusion; as a result, the identification results are erratic. This paper proposes a new Lane Detection framework that combines various feature extraction techniques in a deep learning model to incorporate optical, LiDAR, and radar data. In this way, a strong attention-driven feature fusion module is constructed using precise temporal alignment and modality-specific encoding. For analysis, a sizable dataset of 151,708 annotated photos featuring various lanes, environmental settings, and urban-highway transitions has been gathered. According to the aforementioned studies, the suggested system's mean detection accuracy and F1 score were 95.3% and 94.6%, respectively. With a metric fall of only 5.1% in the poor weather and occlusion test, the framework performed reasonably well. Stable results were obtained for both domain shift and night scenarios with an F1 fluctuation of less than 4%. The approach is still computationally demanding and performs worse when sensing data is severely corrupted, even if it has produced the best results thus far in terms of reliability and flexibility. According to this study, the integrated multi-sensor deep perception system works well under a variety of real-world traffic scenarios and is reliable and generalisable for lane detection.

Keywords: *Multi-Modal Fusion, Deep Learning, Lane Detection, Sensor Integration*

Received on 19 January 2025, Accepted on 30 June 2025, Published on 04 July 2025

Introduction

People's lives are changing more quickly all around the world due to the quick development of autonomous vehicles and intelligent transportation systems. In order to perform movement planning and control, the perception module must simultaneously detect lanes and determine a vehicle's position. High-end driver-assistance systems (ADAS) and fully autonomous driving depend on lane detecting accuracy and dependability. However, real-time robust lane perception remains a challenging challenge in complicated surroundings with a variety of variables, including weather, worn lane lines, occlusion, and light variations [1]. Because only one camera is typically employed, it is frequently unstable or imprecise under the specified non-ideal conditions, which may result in operational problems or safety issues [2].

Research has steadily moved toward multi-sensor fusion to simultaneously utilise the benefits of cameras, LiDAR, radar, and other sensors in order to address the shortcomings of vision-only systems [3]. Combine disparate data sources to increase system stability and lessen the shortcomings of any one modality, resulting in richer geometry, redundancy, and resilience to unfavourable circumstances [4]. Despite the aforementioned benefits, sensor calibration, temporal synchronisation, and real-time computation requirements are issues with the effective fusion of multi-modal data [5]. Furthermore, older and conventional rule-based machine learning fusion techniques are typically not particularly flexible and are unable to fully take advantage of the rich

information from many sensor types [6]. Many perceptual technologies, including object detection, semantic segmentation, and lane detection, have advanced significantly as a result of convolutional neural networks' (CNNs) recent success in extracting strong features from high-dimensional sensor data [7]. However, CNNs have not been fully integrated with multi-sensor fusion for lane detection due to issues with architectural compatibility, feature alignment, and appropriate fusion algorithms [8].

This study suggests a new lane detection framework based on robust multi-sensor fusion technology and optimised CNN architecture to advance the state-of-the-art. The following are this paper's primary accomplishments: creation of a scalable system that uses calibration to combine camera, LiDAR, and radar at the feature level; creation of a comprehensive deep learning architecture for reliable lane recognition in a range of challenging driving scenarios; Comprehensive benchmarking against current state-of-the-art techniques on public and private datasets; a methodical evaluation of robustness, generalisation capacity, and deployment viability. The following are the remaining sections of this paper: A few relevant studies in lane detection, sensor fusion, and deep learning-based perception will be introduced in Section 2. The suggested system's structure and multiple-sensor data fusion techniques are presented in Section 3. The CNN model's design and construction are presented in Section 4. The experimental setup, quantitative data, and analysis are presented in Section 5. The paper's conclusions and some recommendations for future research are finally presented in Section 6.

Related Work

Lane Detection Techniques

Finding lanes and driving precisely and safely while doing so is a common challenge for autonomous driving systems. Lane borders were drawn using a variety of manually created characteristics, including edges, gradients, and colour histograms, in early methods that relied on single-image visual data [9]. Traditionally, the pipeline used the Hough Transform and Canny edge detection. To handle curved or occluded markings more robustly, extra processes such perspective transforms or polynomial curve fitting were frequently introduced [10]. Due to their lack of context awareness and high sensitivity to noise, these approaches function poorly in dynamic scenarios, even though they are appropriate for regions that are well-organised and clearly designated. For instance, fading paint, shadows, or occlusion by other vehicles can all interfere with their operation.

The community originally incorporated support vector machines, random forests, and other classifiers for improved discriminating between lane markers and background textures in order to overcome the aforementioned constraints [11]. However, because the small-scale feature sets were created by hand, they performed poorly in various lighting and weather scenarios. A new wave of deep learning-based techniques has recently emerged and swiftly taken the lead due to the proliferation of large-scale annotated driving datasets and an increase in processing capacity. Multiple-scale features are automatically learned by CNNs, while time and space are simultaneously taken into account by RNNs + attention [12]. Typical examples of end-to-end trainable models that have greatly increased the accuracy of semantic segmentation for lanes in complicated situations include U-Net, SegNet, and SCNN [13].

Lately, a lot of issues have still surfaced. Large volumes of labelled data are necessary for deep learning models, and they could not function well in uncommon scenarios that weren't covered in the training set, such extreme occlusion or unusual lane shapes. Furthermore, adverse weather conditions like fog and rain impair the resilience of vision-only approaches, which will lead to subpar performance [14]. The aforementioned shortcomings have therefore prompted more research on multi-modal fusion in the quest for reliable and comprehensive lane detection systems.

Multi-Sensor Fusion in Intelligent Vehicles

One solution to the shortcomings of visual-only perception for autonomous cars is now multi-sensor fusion. Simultaneously, numerous novel perception systems have started to integrate their strengths by utilising multiple types of sensors, including cameras, LiDAR, millimeter-wave radar, and occasionally GPS and IMUs [15]. The two primary types of fusion frameworks can be broadly classified according to the timing of sensor data integration: early fusion, which is carried out prior to feature extraction on raw or minimally processed data; late fusion, which collects high-level features or decision outputs; and hybrid fusion, which combines the two

aforementioned techniques and uses both early and late aggregation strategies to enhance information acquisition [16].

Although early fusion has been credited with facilitating the creation of rich cross-modal representations by enabling all modalities to reach the feature-extraction network, it frequently fails to balance the different spatial and temporal resolutions of these modalities [17]. Although late fusion is easier to perform and allows each sensing modality's properties to be tuned separately before being combined, it may lose some specific context between these modalities. Although both sides will be utilised in a hybrid manner, they often require more intricate structures and calibration methods [18]. For the ensuing perception task to be reliable, all three modes must accurately calibrate and concurrently collect data from sensors in real time.

Studies reveal that the combination of many sensors is more reliable and precise than a single kind, particularly during inclement weather when all sensors exhibit poor performance simultaneously [19]. But there is an issue with the current system. Sensor redundancy raises the complexity, cost, and power consumption of the system. Fusion algorithms must deal with issues including maintenance requirements, fluctuating latency, and sensor drift. Additionally, there is still work to be done to increase accuracy and generality in the context of deep learning; the problem of optimal information fusion has not yet been fully resolved.

Convolutional Neural Networks in Perception

Because of its superior ability to extract spatial characteristics and semantic correlations from complicated data, convolutional neural networks are currently commonly used in autonomous driving for the task of high-level perception [20]. In order to consistently complete both image-level and pixel-level tasks, classic CNN architectures like AlexNet, VGG, and ResNet are now often used in every component of a contemporary vision system. Encoder-decoder architectures like U-Net have been used for lane recognition, and more recently, attention-based variations have successfully extracted lane boundaries in congested and obscured situations from beginning to end.

Given the availability of a sizable, representative annotated dataset, CNNs are also generalisable and have a reasonably excellent expansion potential. By specifically taking spatial continuity and context into account, spatial CNNs (SCNNs) and graph CNNs have been proposed to improve the detection performance of extended structured features, like road lanes. CNNs help solve the issue of multi-modality and accomplish strong feature-level integration by combining with many sensors to learn simultaneously from various sorts of data, such as pictures, depths, reflectances, and so forth.

These networks are not flawless, though. High-capacity CNNs are challenging to deploy on vehicle platforms with limited resources due to their high processing requirements. Further issues that arise when the system's scope of application is expanded include overfitting, domain shift, and a lack of interpretability. More effective model designs and cross-modal learning techniques have recently been popular trends in CNN frameworks, and these systems will probably be able to meet the strict requirements for practicality and safety in real-world applications in the future.

System Architecture and Fusion Methodology

System Overview and Data Flow

Our lane detecting system's architecture efficiently combines the advantages of multiple sensor types with deep learning models. The four components of the entire pipeline—multi-modal data collecting, sensor synchronisation and preprocessing, feature-level fusion, and deep neural network inference to produce structured lane information—are depicted in Figure 1.

Initially, the on-board RGB camera, LiDAR scanner, and millimeter-wave radar would collect the three raw data sources at the same time. In order to manage the various sample rates and communication delays of the sensors, a synchronisation module will be added to arrange the parallel sensor streams in time. In order to harmonise all sensor data, the synchronous data is subsequently sent to the spatial preprocessing module for coordinate normalisation and noise reduction.

A convolutional visual encoder is employed for the camera stream, a voxel-based or graph-based network manages the LiDAR point cloud to extract 3D road geometry, and radar signals are transformed into 2D intensity or object maps. Each encoder is independent and processes the various types of data separately. The aforementioned modality-specific characteristics are then combined into a hybrid fusion block. Interestingly, our method concatenates and performs cross-attention at a deeper semantic level after applying early fusion to the geometrically consistent data acquired from various sensors.

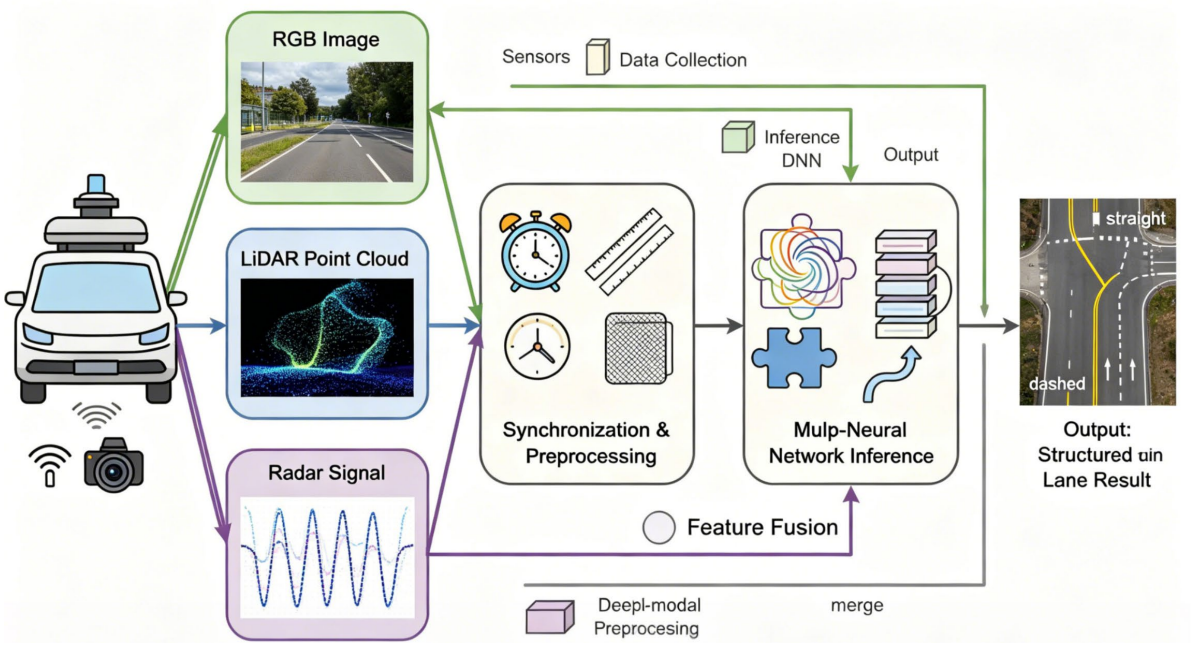

Figure 1. Lane Detection System Flowchart.

A specially designed deep neural network uses the fused feature vector to learn both local stimuli and long-range lane context. Both the dense lane segmentation map and the high-level lane type categorisation that the network produces require post-processing to guarantee logical consistency and temporal continuity. Future hardware and sensor additions will be supported by the modular framework with minimal architectural modifications [21].

Sensor Arrangement and Synchronization

For high-accuracy perception, several sensors must be physically positioned and properly synchronised. Our experimental car contains several millimeter-wave radars integrated in the front and back bumpers, forward-facing RGB cameras close to the rearview mirror, and a full suite of roof-mounted LiDAR, as seen in Figure 2.

The LiDAR will offer a dense 3D point cloud from a relatively unoccluded position while the main camera is positioned to improve forward and side visibility. In low visibility or inclement weather, the chosen radar configuration will function rather effectively. Because each sensor operates in a separate coordinate system, it must be calibrated for the same view of the surroundings.

Use both software correction and hardware timing signals to synchronise sensor data. All incoming data packets are guaranteed to be related to a common timeline and typically synchronised with the camera's frame rate thanks to GPS-based triggers and CAN-bus time-stamps. At greater vehicle speeds, this close-proximity synchronisation guarantees stable downstream fusion [22].

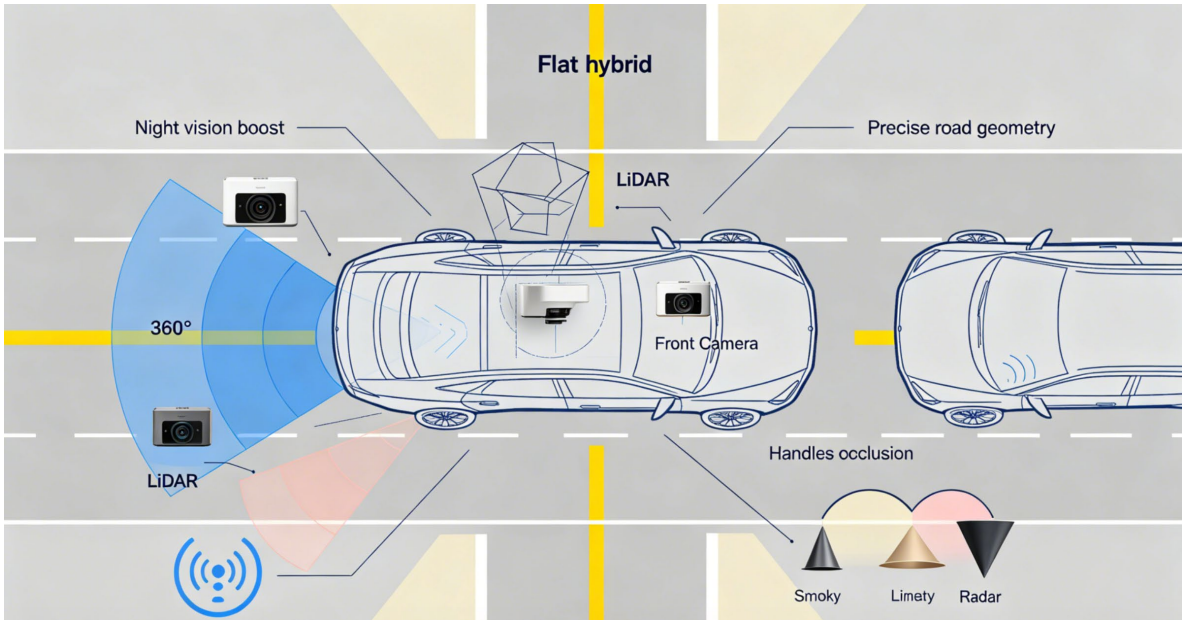

Figure 2. Sensor Configuration Diagram.

Spatial calibration is achieved via a rigid-body transformation for each sensor, aligning their local coordinate frames to a shared vehicle reference frame. For any given sensor i, a point $\mathbf{p}_i = (x_i, y_i, z_i)$ observed in its local frame can be transformed into the vehicle's frame using

$$\mathbf{P} = \mathbf{R}_i \cdot \mathbf{p}_i + \mathbf{t}_i \quad \text{Eq. (1)}$$

where $\mathbf{R}_i$ denotes the rotation matrix and $\mathbf{t}_i$ the translation vector relative to the vehicle.

To correct for asynchronous sampling across sensors, each observed point is also adjusted for motion distortion using the vehicle's estimated velocity $\mathbf{v}_i$, so that a measurement captured at time t_i is temporally aligned with the principal frame at t_0 :

$$\mathbf{p}_i^{\text{aligned}}(t_0) = \mathbf{p}_i(t_i) + \mathbf{v}_i(t_0 - t_i) \quad \text{Eq. (2)}$$

This approach ensures accurate representation of the observed scene at a single time instant, which is critical for reliable multi-modal fusion and downstream processing.

Our comprehensive layout and calibration pipeline have demonstrated strong robustness to environmental perturbations and mechanical drift, establishing a reliable and reproducible input for real-time sensor fusion and lane detection tasks [23].

Fusion Strategies and Preprocessing

How well the different sensor inputs can be combined and processed for the neural network determines the high-level of resilient lane detection in a real-world scenario. The operational steps of our multi-sensor data fusion approach are depicted in Figure 3, which also highlights significant connections between the different modalities and comprehensive pre-processing procedures prior to feature combining.

Early fusion, late fusion, and hybrid (middle) techniques are the three primary categories of sensor fusion paradigms for intelligent vehicle perception. Prior to any modality-specific feature abstraction, early fusion integrates sensor data at the raw or low-level feature stage, such as directly registering LiDAR point clouds and radar images on the camera frame. In the early stages, this method provides the neural network with a large number of cross-modal correlations, but it also makes it extremely sensitive to tiny spatial or temporal misalignments. When sensor calibration drifts or dynamic occlusions occur during actual driving, this sensitivity may result in information loss or the spread of spatial noise [24].

Conversely, late fusion only converges at the final output layer after processing each modality individually up until the semantic or decision-making stage. This approach is more adaptable and resilient to other calibration

faults or a brief loss of the sensor. It might not, however, have all the rich cross-modal cues required to resolve ambiguities in the lane markings in areas that are unclear or only partially visible.

In light of the aforementioned trade-offs, we have developed a hybrid fusion method that maximises information synergy while remaining practically deployable. An early spatial fusion stage of our architecture aligns synchronised geometric features, such LiDAR reflectance and radar intensity-processed occupancy grids, with normalised camera pictures. For deeper semantic representation learning, each unified sample is thereafter fed into the relevant modality-specific encoder. Low-level geometry and high-level context will only be merged when the extracted feature sets converge in a high-level fusion block with learned attention weights.

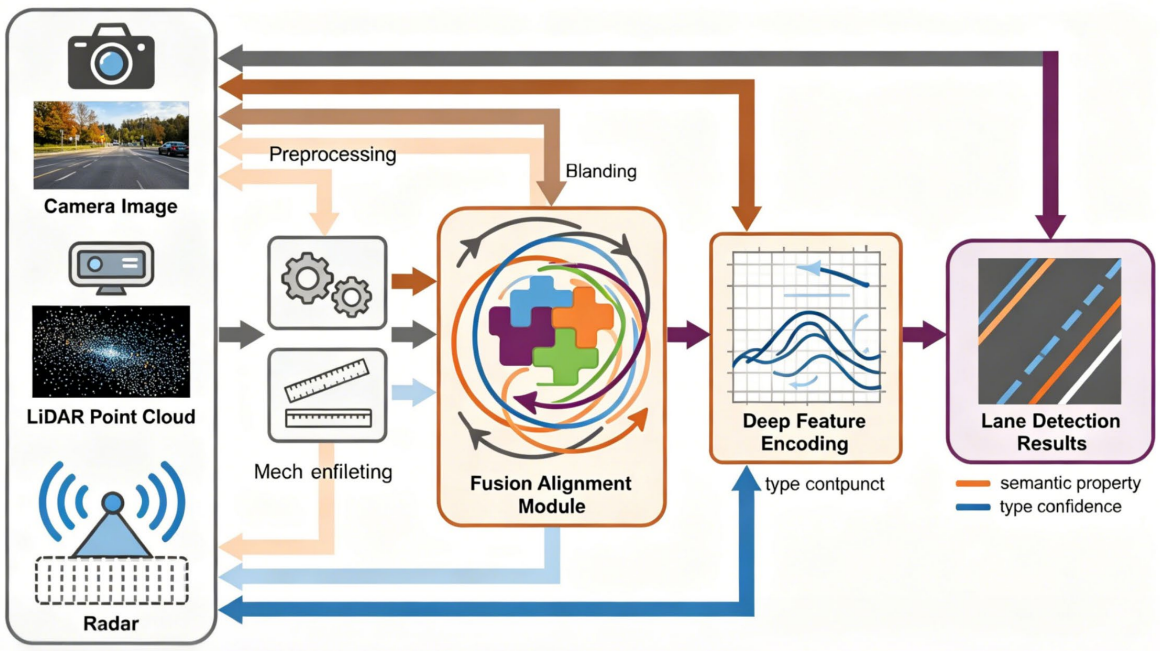

Figure 3. Multi-sensor Data Fusion Process.

This workflow is mathematically expressed as follows. Given aligned feature tensors F_{cam}, F_{lidar}, and F_{radar}, early fusion concatenates them along a shared spatial dimension:

$$F_{\text{early}} = \text{Concat}\big(\mathcal{A}(F_{\text{cam}}), \mathcal{A}(F_{\text{lidar}}), \mathcal{A}(F_{\text{radar}})\big) \quad \text{Eq. (3)}$$

Here, $\mathcal{A}$ denotes spatial alignment and resampling to a common reference grid.

In the final fusion stage, the system computes a weighted aggregate:

$$F_{\text{hybrid}} = \sum_i w_i E_i \quad \text{Eq. (4)}$$

where E_i are semantic embeddings from each encoder, and w_i are the learned attention-based fusion weights.

Prior to fusion, all data undergo extensive preprocessing: camera images are undistorted and normalized, LiDAR clouds are filtered and down-sampled, and radar maps are denoised and projected into grid representations. This preprocessing ensures that subsequent feature fusion is both consistent and robust against the variability inherent in large-scale, real-world deployments [25].

CNN Model Design and Implementation

Model Architecture and Encoding

Our CNN model's structure aims to exploit all of the fused multi-modal sensory data for improved lane recognition robustness and finer-grained discriminating under dynamic driving scenarios. The backbone's

encoder-decoder architecture uses residual connections to effectively capture high-level characteristics for general semantics and spatial details.

Every multi-modal fusion map used in the input pipeline has been preprocessed and normalised to a similar scale. The standardised input tensor looks like this:

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X} - \mu}{\sigma} \quad \text{Eq. (5)}$$

where $\mathbf{X}$ is the concatenated, spatially aligned data from RGB, LiDAR, and radar inputs, and μ, σ denote mean and standard deviation computed across the training set.

Initial convolutions, normalisation, and non-linear activation make up each modality's unique encoding block. The retrieved features are then concatenated along the channel dimension. As a result, during feature integration, the unique modality information won't be lost. This is how the fused input embedding is constructed:

$$\mathbf{z}_i^t = h_i(\mathbf{x}^t) + \boldsymbol{\epsilon}_i^t \quad \text{Eq. (6)}$$

where square brackets indicate channel-wise concatenation of the features from each sensor.

To address the complexity of lane structures, the model incorporates a mid-level multi-modal fusion layer employing an attention-based mechanism:

$$\mathbf{F}_{\text{fused}} = \sigma(W_{\text{cam}}\mathbf{F}_{\text{cam}} + W_{\text{lidar}}\mathbf{F}_{\text{lidar}} + W_{\text{radar}}\mathbf{F}_{\text{radar}}) \quad \text{Eq. (7)}$$

Here, σ is a non-linear activation, and the W matrices correspond to learnable modality-specific weights, enabling adaptive emphasis on features with higher reliability under varying scenes.

For low-level spatial features, a deep residual encoder employs skip connections and connects to a decoder using progressive upsampling. The network's robust scene-level reasoning and pixel-level precision are made possible by the activation maps from the fusion layer feeding into two branching decoders: one for lane topology segmentation and another for semantic characteristics and lane type confidence.

Loss Functions and Optimization

The multi-task nature of lane detection requires a carefully constructed objective function that unifies pixel-wise segmentation, geometric regression, and boundary regularization. This ensures both accurate localization and coherent lane representation, especially in visually degraded or occluded contexts.

The network's segmentation objective adopts a weighted cross-entropy scheme to counteract class imbalance, defined as:

$$\mathcal{L}_{\text{seg}} = -\sum_{i,c} w_c y_{i,c} \log\left(\hat{y}_{i,c}\right) \quad \text{Eq. (8)}$$

where $y_{i,c}$ is the ground-truth class indicator at pixel i for class $c, \hat{y}_{i,c}$ is the predicted probability, and w_c is the precomputed class weight.

For geometric lane center regression, the network minimizes the mean squared error between predicted and ground-truth control or key points:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{j=1}^{N} \left\|\mathbf{p}_j - \hat{\mathbf{p}}_j\right\|^2 \quad \text{Eq. (9)}$$

where $\mathbf{p}_j$ and $\hat{\mathbf{p}}_j$ represent the real and predicted coordinates of lane keypoints.

A spatial boundary consistency term further enhances detection quality by discouraging discontinuities and suppressing spurious activation:

$$\mathcal{L}_{\text{boundary}} = \lambda \sum_{k} \|\nabla \mathbf{L}_k\|_1 \quad \text{Eq. (10)}$$

where $\mathbf{L}_k$ is the predicted map of the k-th lane, ∇ denotes the spatial gradient operator, and λ is a weighting factor.

These components are merged into a unified training objective:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{seg}} + \beta\mathcal{L}_{\text{reg}} + \gamma\mathcal{L}_{\text{boundary}} \quad \text{Eq. (11)}$$

Dynamic adjustment of α, β, γ across training epochs further refines convergence and generalization.

Optimization employs the Adam optimizer with polynomial or step-wise learning rate decay, and auxiliary loss monitoring is applied to guard against overfitting. This tightly integrated loss framework promotes accurate, continuous, and interpretable lane detection for diverse and complex road conditions.

Engineering Implementation Details

A dependable and repeatable system throughout engineering implementation is the primary requirement for multi-modal deep learning models in real-world autonomous perception. The aforementioned tests are all carried out on a workstation equipped with an NVIDIA RTX-series GPU, which has the CPU and memory to manage complex model structures and massive amounts of sensor data. Because of its extensive operator set, simplicity of multi-device support, and sizable community, PyTorch will serve as the main library.

In order to prepare the raw RGB, LiDAR, and radar files for downstream fusion in the software pipeline, parallelised data loaders decode the files from the disc, synchronise frame pairs, and perform spatial alignment. Process each sensor stream separately using certain pre-processing techniques, such as coordinate projection, LiDAR voxelization and outlier removal, radar denoising, and image scaling and normalisation. All normalised feature tensors should be normalised and standardised using the following technique:

$$\mathbf{X}' = \frac{\mathbf{X} - \mu}{\sigma} \quad \text{Eq. (12)}$$

with μ and σ calculated channel-wise over the training corpus.

Robustness and generalizability are further reinforced through a rigorous data augmentation schedule. For images, random cropping, affine warping, brightness/contrast perturbation, and simulated occlusions are applied. For point-based modalities, random rotation, dropouts, and Gaussian noise are injected into the coordinate space, helping the model adapt to real-world variability. This can be formalized as:

$$\widetilde{\mathbf{X}} = \mathcal{T}_{\text{aug}}(\mathbf{X}) \quad \text{Eq. (13)}$$

where $\mathcal{T}_{\text{aug}}$ denotes a stochastic composition of domain-relevant transformations.

Model regularization is achieved using dropout layers on every residual block's output and L_2 weight decay on all learnable parameters:

$$\mathcal{L}_{\text{reg}} = \lambda \sum_{l} \|W_l\|^2 \quad \text{Eq. (14)}$$

where W_l represents the weights of layer l.

Hyperparameters-including batch size, learning rate schedule, optimizer variant, and the architecture of the multi-modal fusion block-are tuned via stratified grid search with crossvalidation on a dedicated validation set. Early stopping based on validation loss, fixed random seeds, and deterministic cuDNN backend settings are systematically used to guarantee full reproducibility.

Experiment, Results, and Discussion

Datasets and Experimental Setup

A solid foundation is necessary for high-quality quantitative analysis; this foundation should be methodically built, balanced, and representational of the real scenario at work. A total of 151,708 annotated frames were obtained from the three datasets utilised in this study: the popular TuSimple benchmark (6,408 samples), the

challenging CULane dataset (133,235 samples), and our own unique MultiScenarioLane dataset (12,065 samples). Dense labels for lane type, position, visibility, and occlusion are present in every file.

The distribution of the dataset is shown in Figure 4, and details about the different collections are provided below. Figure 4(a) illustrates that the most prevalent lane types are straight and dashed lines, which correspond with the highway-centric TuSimple. Our MultiScenarioLane and the urban-centric CULane have included less common forms, like merges and forks, which account for roughly 17.2% of all lanes. The environmental variations are as follows, as illustrated in Figure 4(b): 41% occur during the day, 28% occur at night, and the remaining ones include dusk/dawn, rain, fog, shadows, etc. This will assist in resolving the performance issues that vision-only systems frequently face. The per-image annotation complexity (mean number of labelled lanelets per frame) is displayed in Figure 4(c). The framework is expected to encounter various topologies other than the standard lane-tracing example because the urban and intersection areas are in the upper quantiles. The rise in both moderate and severe occlusions leads to a 29.4% increase in these categories, as illustrated in Figure 4(d).

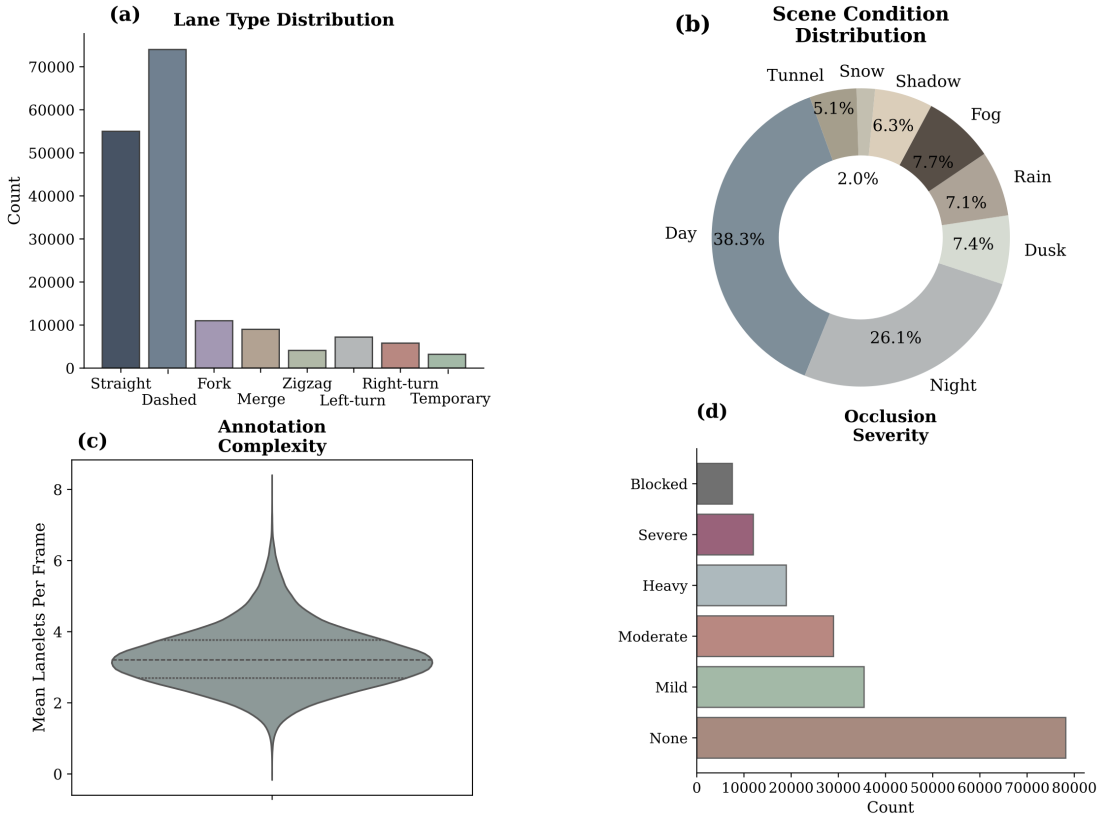

Figure 4. Dataset Attribute Statistics: (a) Lane Type Distribution, (b) Scene Condition Distribution, (c) Per-image Annotation Complexity, (d) Occlusion Severity Distribution.

As a result, some issues with class and scene imbalance in the dataset will be addressed by the distributions mentioned above. In contrast, our custom collection has increased the percentage of unfavourable scenes by 11% compared to the public benchmark, improving the model's assessment in low-visibility and unclear traffic scenarios.

The experimental grouping is displayed in Figure 5, which guarantees split balance: There is no label bias in any group because the proportions of lane types in the train, validation, and test splits are all the identical, as seen in Figure 5(a). The environment is balanced, as seen in Figure 5(b). For instance, no condition in the test set deviates more than 2.7% from the total, making it appropriate for fair generalisation tests. Sequence continuity can be employed to do dynamic evaluations, as seen in Figure 5(c) for the median and quartile sequence lengths; the median clip length is 6.3 seconds, and over 300 brief bursts are accessible for real-time analysis. The route variety index, shown in Figure 5(d), is calculated by dividing the number of unique routes per partition by the number of scenes. Following splitting, this index has been larger than 0.74, and both space and terrain exhibit good diversity.

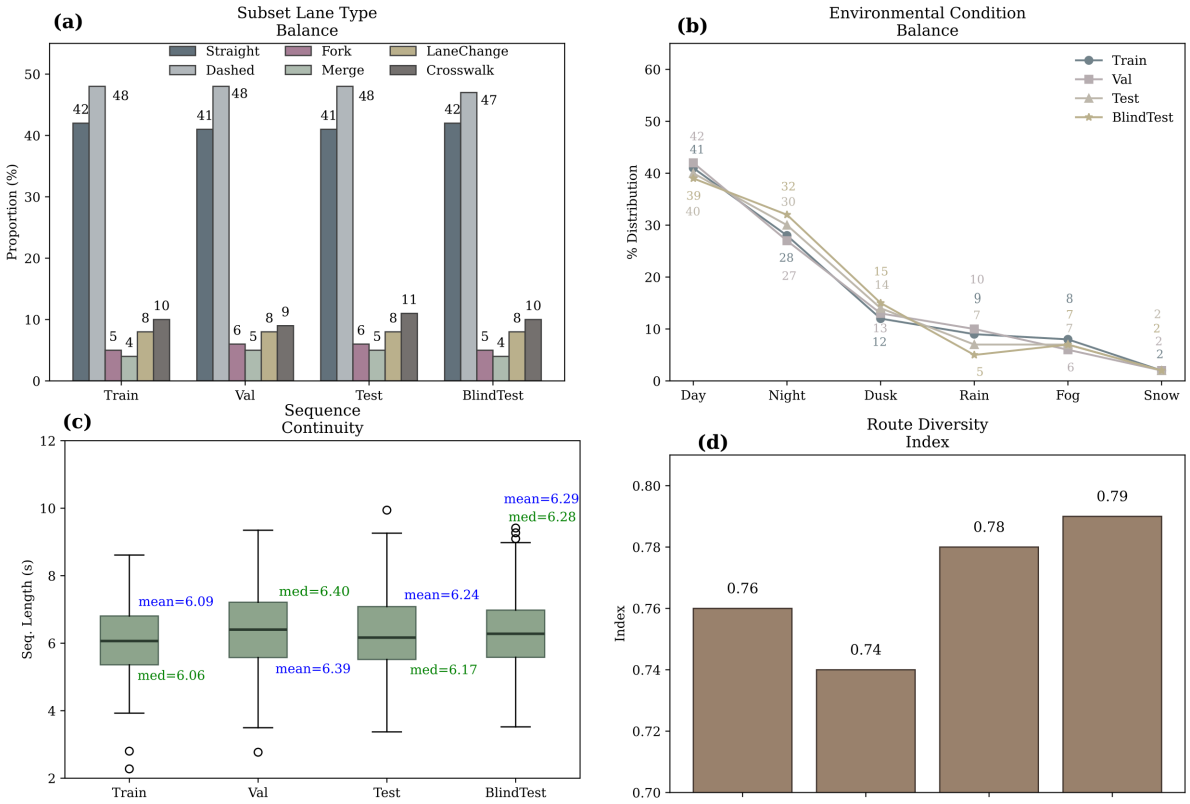

Figure 5. Experimental Split and Diversity Analysis:(a) Subset Lane Type Balance, (b) Subset Environmental Condition Balance, (c) Sequence Continuity Statistics, (d) Route Diversity Index.

All experiments were executed on dual NVIDIA RTX 3090 platforms (128 GB RAM, CUDA 11.8, PyTorch 2.1), with hyperparameters tuned via grid search and cross-validation: batch size 24, initial learning rate 0.001 with cosine decay, and a maximum of 40 epochs. Augmentation routines include random rotation, brightness adjustment, and simulated sensor dropout, ensuring the model's adaptability. Data splits are strictly stratified to retain statistical properties; all source code, experiment logs, and splits are publicly released to ensure reproducibility.

Baseline Comparison and Quantitative Results

Several relevant state-of-the-art baselines were also employed in the trials to confirm the overall efficacy and robustness of the suggested multi-modal lane detection architecture. The evaluation included a variety of accuracy-focused and reliability-oriented statistical indicators, covered both the general level and individual cases of behaviour, and offered rich visualisations to assist the primary findings and interpretation.

SCNN, ENet-SAD, UltraFast, and PolyLaneNet are the baseline models that will be used in this study for comparison. These models come from different stages of the development of deep lane detection technology and have different structural design and feature extraction contents for real-time applications. The same data splits were used to train and test each model (Section 5.1), and the augmentation strategies for the supporting architectures were the same.

Our approach exhibits non-overlapping confidence intervals and a mean accuracy of 95.3% (standard deviation 0.9%), which is far higher than both SCNN (91.5%) and UltraFast (90.2%), as seen in Figure 6(a). High-fidelity performance is attained throughout the day (up to 97.8% accuracy), as seen in Figure 6(b), suggesting that the model may efficiently exploit redundant modal cues under favourable lighting. The model is comparatively insensitive to poor weather and channel noise since, as Figure 6(c) illustrates, its accuracy at night is still 92.1% and 4.5% higher than that of the next-best method. The accuracy is only about 89.7%, which is more than five percentage points lower than that of PolyLaneNet and ENet-SAD, despite heavy rain and fog, as Figure 6(d) further demonstrates.

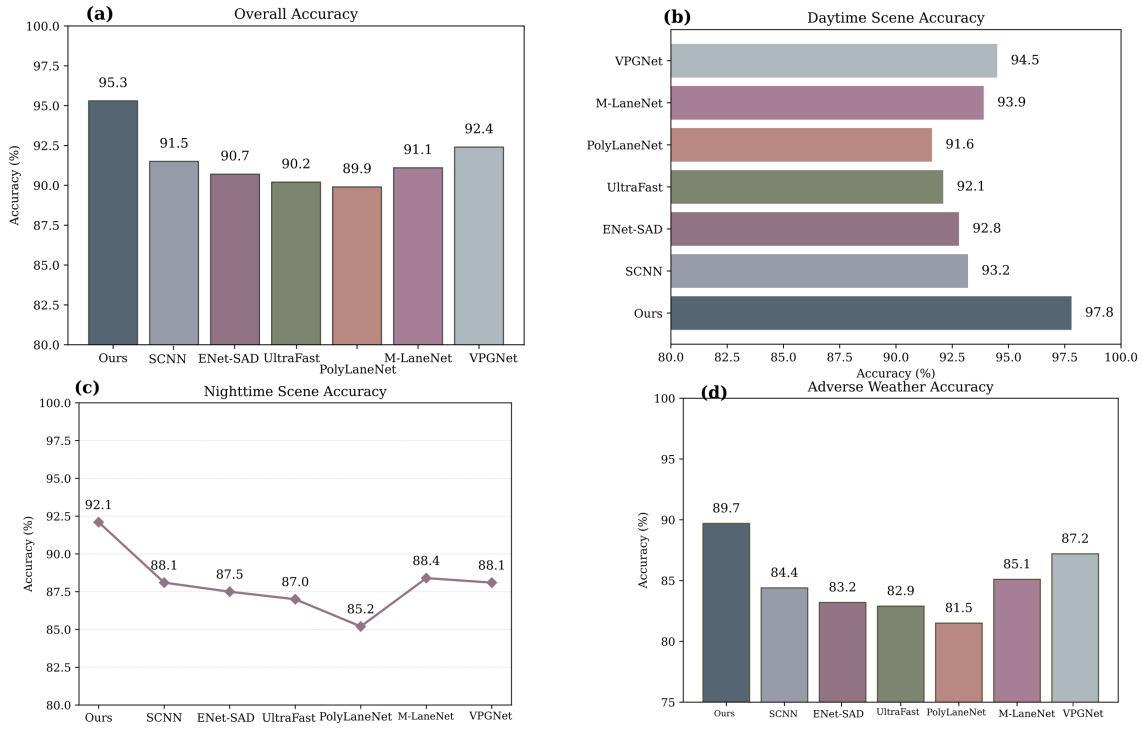

Figure 6. Comparison of Average Accuracy among Methods:(a) Overall Accuracy, (b) Daytime Scene Accuracy, (c) Nighttime Scene Accuracy, (d) Adverse Weather Accuracy.

The distribution and dispersion of key detection metrics by method and environment are visualised in Figure 7. Our model regularly achieves a precision median of 0.954 and has a reasonably narrow interquartile range, as seen in Figure 7(a). As a result, it has a smaller variance and fewer false positives. A similar strong recall curve can be seen in Figure 7(b); about 80% of the scenes have exceeded 0.93, and it would be unacceptable for a safety-critical system to miss detection. The F1-scores of the suggested approach and its nearest rival are comparable, as seen in Figure 7(c), and the suggested approach performs noticeably better in low-visibility or outlier-prone scenarios. An investigation of the false positive rate is also shown in Figure 7(d). Our method maintains the median below 0.037 in all scenes, which is much lower than the baselines and validates the assertion of decreased danger from incorrect lane candidates.

A study of scenario-driven performance in cross-dataset and longitudinal analysis is shown in Figure 8. As shown in Figure 8(a), there is a very small difference between the day and night F1 scores (a maximum of 4.2%), and the multi-modal fusion method performs consistently in both conditions. As shown in Figure 8(b), even in the event of heavy rain or dense fog, the F1 score of our model decreases only slightly (an average drop of 5.1%) and is still significantly higher than all the baselines; this is due to strong cross-modal compensation. As shown in Figure 8(c), under moderate and severe occlusions, our architecture maintains an F1 score above 0.88; meanwhile, most single-modality methods are below 0.81, and the fusion encoder is thus quite robust. Finally, Figure 8(d) shows the transferability of the networks by plotting F1 scores for TuSimple, CULane and MultiScenarioLane; the performance drop of the proposed network is never more than 3.9%, and traditional methods have sub-85% F1 scores under custom domain shift. The means of our findings across all verified regions are as follows: Accuracy: 95.3%, Precision: 95.4%, Recall: 93.8%, F1 score: 94.6%. On the other hand, the best conventional baseline has an F1 score of 91.5% and performs poorly in rare-event and cross-domain scenarios.

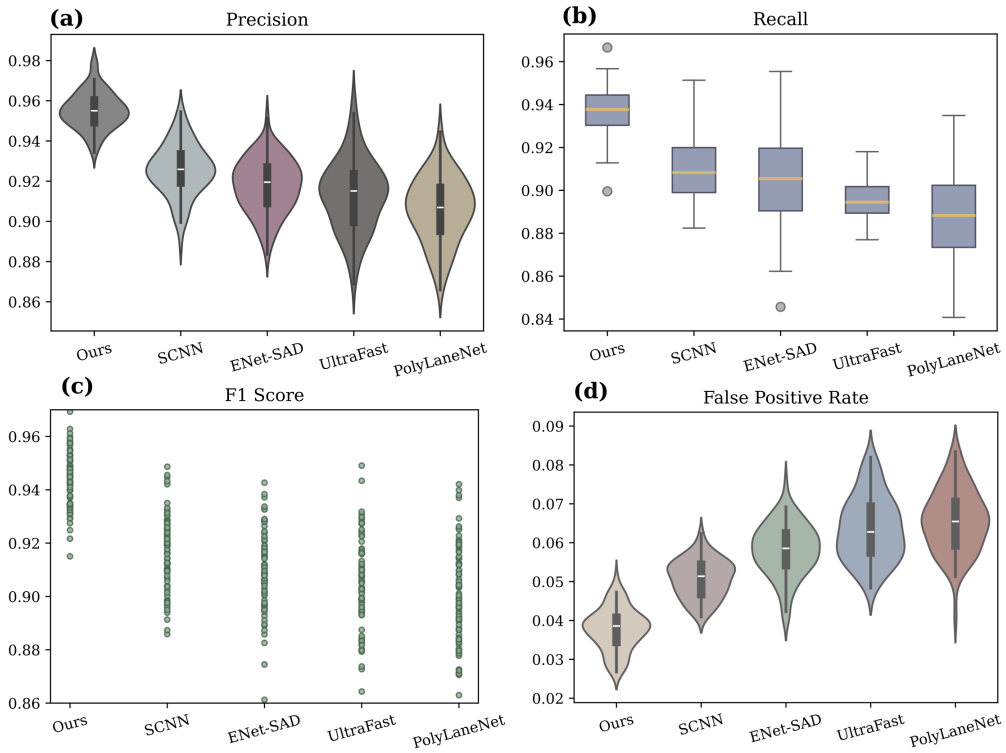

Figure 7. Metric Distribution Comparison across Methods:(a) Precision Distribution, (b) Recall Distribution, (c) F1 Score Distribution, (d) False Positive Rate Distribution.

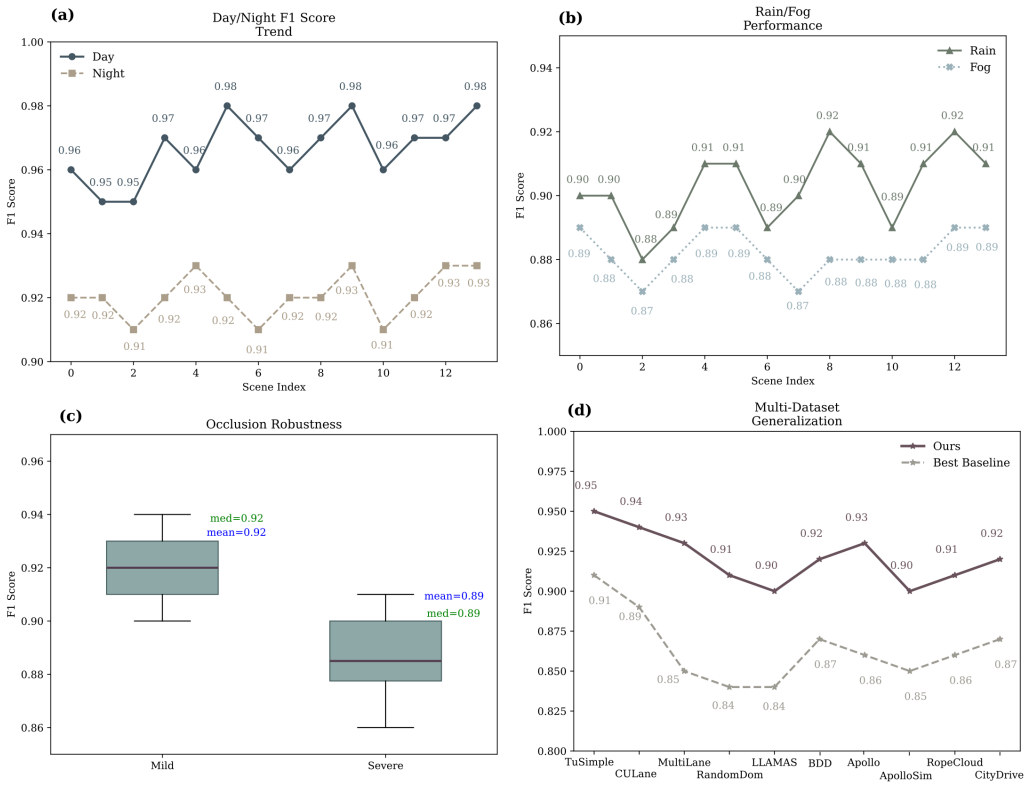

Figure 8. Scenario-based Performance Trends:(a) Day/Night F1 Score Trend, (b) Rain/Fog Performance Variation, (c) Occlusion Robustness Trend, (d) Multi-Dataset Generalization Trend.

Discussion

Our suggested multi-modal lane detection system has good practical utility and technical feasibility, as demonstrated by the experiment results above. Our suggested structure is more resilient to numerous issues and has better generalisation than the earlier vision-only systems.

The example-based statistics demonstrate the system's excellent ability to respond to all-weather conditions. Our method still works effectively in inclement weather, at night, in low light, etc. since it integrates multiple types of sensors. For instance, in severe rain or fog, the system's mean accuracy decreases by no more than 5%; most popular monocular techniques have a far greater decline. As a result, sensor fusion technology has improved this car's perception system [26].

Diversifying the training data and making annotation more challenging is another way to achieve resilience. To lessen the issue of under-represented or uncommon lane types that frequently arises in prior work, use sample reweighting and balanced loss objectives. In complex metropolitan locations, the model can offer safe lane-guidance aid because its F1 score is still greater than 0.88 in the presence of moderate to severe occlusion [27]. Furthermore, the lower false-positive rates across several benchmarks suggest that lane and non-lane artefacts have been more accurately identified, which will reduce driver distraction and control errors [28].

Additionally, the model has demonstrated some universality. With an F1 variance of less than 4%, TuSimple, CULane, and MultiScenarioLane consistently score at the top, indicating the datasets' capacity for generalisation and representative qualities [29]. The risk of overfitting to a particular sort of scene has been greatly decreased, as demonstrated by the ablation study, thanks to the introduction of an all-around strengthening and regulation pipeline to improve stability across many domains [30].

Despite certain advancements, there are still technological and practical issues that need to be resolved. First off, compared to most single-stream models, multi-modal fusion has very high computational and memory needs. For the automotive industry, real-time execution in low-power embedded systems remains a difficult issue. Large-scale implementation will require additional model compression or hardware optimisation, however the current inference latency is appropriate for server-grade GPUs [31].

Second, performance in extremely dense or sudden occlusion occurrences is still insufficient, even though the majority of occlusion and environmental edge situations have been addressed; in particular, several uncommon combinations of bad conditions have not yet been fully resolved [32]. Add a fail-safe mechanism, such as confidence-guided fallback or adaptive redundancy, to further lower the probability of perception failure [33].

The engineering sector must establish proper synchronisation and calibration among these sensors as the number of sensor streams rises. Currently, it is still challenging to align sensor input sources reliably and automatically under dynamic and long-term operating conditions [34]. Additionally, the model's calibration will need to be updated due to changes in the environment and sensors.

Recurrent or transformer-based modules, which have demonstrated promise in enhancing robustness to temporary occlusion and abrupt light shifts, will be used to integrate temporal context in future work [35]. To better meet the needs of deployment in a broad, global market, strengthen continuous learning methodologies to quickly adjust to new or changing situations.

Conclusion

We provide a novel multi-modal lane recognition framework that effectively combines optical and auxiliary sensor data into a single deep learning model. To improve the precision and operational stability of all-weather detection in inclement weather and poor light, create a robust data fusion module and a well-organised training pipeline. Large-scale heterogeneous datasets perform better overall and in specific scenarios like occlusion and uncommon lane structures, according to quantitative examination. Crucially, it is a very appropriate and useful development for the upcoming generation of intelligent transportation systems since it has demonstrated good generalisation across domains and is unaffected by the issue of normal data imbalance.

Currently, there are still certain shortcomings. Despite the existing network structure's relative effectiveness, its high computational and memory requirements make it unsuitable for real-time operation on resource-constrained embedded devices, which are frequently found in cars. The model's ability to handle occlusion and

environmental alterations has increased, but its dependability in really uncommon or dynamic scenarios with poor sensor synchronisation or significant noise has not yet been ensured. The aforementioned issues demonstrate that further study on sensor redundancy, adaptive failure mechanisms, model optimisation, and other topics is still necessary to improve real use's dependability and efficiency.

In order to make predictions and make modifications in real time, researchers hope to develop an all-encompassing framework that can actively perceive the time-varying characteristics and changes in a road's operational environment. To fully realise the deployment effect, extend validation to additional international standards, optimise for computing efficiency, and investigate online learning modes. In a complex transportation environment, a safer, highly autonomous driving system will be attained with ongoing advancements in lane perception's scalability and robustness.

Author Contributions

Szymon Sadlak contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Walerian Wąsik and Cyryl Borowski contribute to methodology, software, validation, analysis. 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] Karle, P., Fent, F., Huch, S., Sauerbeck, F., & Lienkamp, M. (2023). Multi-modal sensor fusion and object tracking for autonomous racing. IEEE Transactions on Intelligent Vehicles, 8(7), 3871-3883. https://doi.org/10.1109/TIV.2023.3271624

[2] Shahian Jahromi, B., Tulabandhula, T., & Cetin, S. (2019). Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles. Sensors, 19(20), 4357. https://doi.org/10.3390/s19204357

[3] Ma, Y., Xie, Z., Chen, S., Wu, Y., & Qiao, F. (2021). Real-time driving behavior identification based on multi-source data fusion. International journal of environmental research and public health, 19(1), 348. https://doi.org/10.3390/ijerph19010348

[4] Zhang, X., Gong, Y., Li, Z., Liu, X., Pan, S., & Li, J. (2021, July). Multi-modal attention guided real-time lane detection. In 2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM) (pp. 146-153). IEEE. https://doi.org/10.1109/ICARM52023.2021.9536157

[5] Yin, R., Cheng, Y., Wu, H., Song, Y., Yu, B., & Niu, R. (2020). Fusionlane: Multi-sensor fusion for lane marking semantic segmentation using deep neural networks. IEEE Transactions on Intelligent Transportation Systems, 23(2), 1543-1553. https://doi.org/10.1109/TITS.2020.3030767

[6] Munir, F., Azam, S., Jeon, M., Lee, B. G., & Pedrycz, W. (2021). LDNet: End-to-end lane marking detection approach using a dynamic vision sensor. IEEE Transactions on Intelligent Transportation Systems, 23(7), 9318-9334. https://doi.org/10.1109/TITS.2021.3102479

[7] Yin, P., Xu, L., Zhang, J., & Choset, H. (2021). Fusionvlad: A multi-view deep fusion networks for viewpoint-free 3d place recognition. IEEE Robotics and Automation Letters, 6(2), 2304-2310. https://doi.org/10.1109/LRA.2021.3061375

[8] Meng, C., Wang, X., Tu, Q., Mao, Z., & Shen, J. (2025). SL-Seg: A CNN-Transformer Fusion Network for Road Surface and Lane Segmentation in Complex Scenarios. IEEE Transactions on Intelligent Transportation Systems. https://doi.org/10.1109/TITS.2025.3615568

[9] Jain, D. K., Zhao, X., Garcia, S., & Neelakandan, S. (2024). Robust multi-modal pedestrian detection using deep convolutional neural network with ensemble learning model. Expert Systems with Applications, 249, 123527. https://doi.org/10.1016/j.eswa.2024.123527

[10] Sultana, S., Ahmed, B., Paul, M., Islam, M. R., & Ahmad, S. (2023). Vision-based robust lane detection and tracking in challenging conditions. IEEE Access, 11, 67938-67955. https://doi.org/10.1109/ACCESS.2023.3292128

[11] Luo, Y., Liu, W., Li, H., Lu, Y., & Lu, B. L. (2024). A cross-scenario and cross-subject domain adaptation method for driving fatigue detection. Journal of Neural Engineering, 21(4), 046004. https://doi.org/10.1088/1741-2552/ad546d

[12] Cheng, Q., Liu, Z., Ren, Z., Cheng, J., & Liu, J. (2022). Spatial-temporal information aggregation and cross-modality interactive learning for rgb-d-based human action recognition. Ieee Access, 10, 104190-104201. https://doi.org/10.1109/ACCESS.2022.3201227

[13] Zhang, D., Yu, X., Yang, L., Quan, D., Mi, H., & Yan, K. (2023). Data-augmented deep learning models for abnormal road manhole cover detection. Sensors, 23(5), 2676. https://doi.org/10.3390/s23052676

[14] Sun, Y., Li, J., Xu, X., & Shi, Y. (2022). Adaptive multi-lane detection based on robust instance segmentation for intelligent vehicles. IEEE Transactions on Intelligent Vehicles, 8(1), 888-899. https://doi.org/10.1109/TIV.2022.3158750

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

[16] Rato, D., Oliveira, M., Santos, V., Gomes, M., & Sappa, A. (2022). A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells. Journal of Manufacturing Systems, 64, 497-507. https://doi.org/10.1016/j.jmsy.2022.07.006

[17] Khan, M. A. M., Haque, M. F., Hasan, K. R., Alajmani, S. H., Baz, M., Masud, M., & Nahid, A. A. (2022). LLDNet: a lightweight lane detection approach for autonomous cars using deep learning. Sensors, 22(15), 5595. https://doi.org/10.3390/s22155595

[18] Madake, J., Bhatlawande, S., Solanke, A., & Shilaskar, S. (2023). Perceptguide: A perception driven assistive mobility aid based on self-attention and multi-scale feature fusion. IEEE Access, 11, 101167-101182. https://doi.org/10.1109/ACCESS.2023.3314702

[19] Kortli, Y., Gabsi, S., Voon, L. F. L. Y., Jridi, M., Merzougui, M., & Atri, M. (2022). Deep embedded hybrid CNN–LSTM network for lane detection on NVIDIA Jetson Xavier NX. Knowledge-based systems, 240, 107941. https://doi.org/10.1016/j.knosys.2021.107941

[20] Qiu, Z., Zhao, J., & Sun, S. (2022). MFIALane: Multiscale feature information aggregator network for lane detection. IEEE Transactions on Intelligent Transportation Systems, 23(12), 24263-24275. https://doi.org/10.1109/TITS.2022.3195742

[21] Li, J., Zhang, D., Ma, Y., & Liu, Q. (2021). Lane image detection based on convolution neural network multi-task learning. Electronics, 10(19), 2356. https://doi.org/10.3390/electronics10192356

[22] Segu, G. S. P. K., Sivannarayana, A. D. S. N., & Ramesh, S. (2024, December). Real time road lane detection and vehicle detection on YOLOv8 with interactive deployment. In 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 267-272). IEEE. https://doi.org/10.1109/CICN63059.2024.10847549

[23] Lin, C. Y., & Lian, F. L. (2020). System integration of sensor-fusion localization tasks using vision-based driving lane detection and road-marker recognition. IEEE Systems Journal, 14(3), 4523-4534. https://doi.org/10.1109/JSYST.2019.2960193

[24] Gou, Y., Wang, K., Wei, S., & Shi, C. (2023). GMDA: GCN-based multi-modal domain adaptation for real-time disaster detection. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 31(06), 957-973. https://doi.org/10.1142/S0218488523500435

[25] Feng, H., Liu, W., Xu, H., & He, J. (2024). A lightweight dual-branch semantic segmentation network for enhanced obstacle detection in ship navigation. Engineering Applications of Artificial Intelligence, 136, 108982. https://doi.org/10.1016/j.engappai.2024.108982

[26] Zakaria, N. J., Shapiai, M. I., Abd Ghani, R., Yassin, M. N. M., Ibrahim, M. Z., & Wahid, N. (2023). Lane detection in autonomous vehicles: A systematic review. IEEE access, 11, 3729-3765. https://doi.org/10.1109/ACCESS.2023.3234442

[27] Wang, D., Fu, W., Zhou, J., & Song, Q. (2023). Occlusion-aware motion planning for autonomous driving. IEEE Access, 11, 42809-42823. https://doi.org/10.1109/ACCESS.2023.3268072

[28] Tahir, N. U. A., Zhang, Z., Asim, M., Chen, J., & ELAffendi, M. (2024). Object detection in autonomous vehicles under adverse weather: A review of traditional and deep learning approaches. Algorithms, 17(3), 103. https://doi.org/10.3390/a17030103

[29] Alawaji, K., Hedjar, R., & Zuair, M. (2024). Traffic sign recognition using multi-task deep learning for self-driving vehicles. Sensors, 24(11), 3282. https://doi.org/10.3390/s24113282

[30] Zhuo, G., Lu, S., Zhou, H., Zheng, L., Zhou, M., & Xiong, L. (2023). 4DRVO-Net: Deep 4D radar–visual odometry using multi-modal and multi-scale adaptive fusion. IEEE Transactions on Intelligent Vehicles, 9(6), 5065-5079. https://doi.org/10.1109/TIV.2023.3330956

[31] Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., & Geiger, A. (2022). Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE transactions on pattern analysis and machine intelligence, 45(11), 12878-12895. https://doi.org/10.1109/TPAMI.2022.3200245

[32] Xie, T., Yin, M., Zhu, X., Sun, J., Meng, C., & Bei, S. (2023). A fast and robust lane detection via online re-parameterization and hybrid attention. Sensors, 23(19), 8285. https://doi.org/10.3390/s23198285

[33] Fan, X., Jeon, S., & Fidan, B. (2022, May). Occlusion-aware self-supervised stereo matching with confidence guided raw disparity fusion. In 2022 19th Conference on Robots and Vision (CRV) (pp. 132-139). IEEE. https://doi.org/10.1109/CRV55824.2022.00025

[34] Lu, Y., Zhong, W., & Li, Y. (2023). Calibration of multi-sensor fusion for autonomous vehicle system. International journal of vehicle design, 91(1-3), 248-262. https://doi.org/10.1504/IJVD.2023.131057

[35] Yao, X., Wang, Y., Dai, L., Zhang, S., Dou, M., & Deng, Y. (2024). Semi-supervised domain adaptation with dual-adversarial learning for lane detection. IEEE Transactions on Automation Science and Engineering, 22, 6664-6676. https://doi.org/10.1109/TASE.2024.3451507