10.48047/jocaaa.2025.34.11.37

Advancements in Real-Time Edge Deployment for Autonomous Truck Perception Systems

Sowmiya Narayanan Govindaraj
University of Minnesota, Twin Cities, USA

Abstract

This article discusses the development and deployment of multi-modal perception systems in autonomous vehicles, the general issue of finding a trade-off between computational efficiency and the reliability of perception in the edge computing setting. Autonomous vehicle perception systems have to handle a variety of sensor data, such as LiDAR, camera, radar, and GPS data, with milliseconds latency requirements and severe power and thermal constraints. There have been fundamental innovations in numerous areas to bridge the gap between research models and production-ready systems: the use of modular pipeline architectures that support parallel processing of various sensor modalities; the use of sophisticated model optimization algorithms such as structured pruning, quantization-aware learning, and knowledge distillation; the use of advanced sensor fusion schemes that can process asynchronous data streams of varying refresh rates; the use of environmental robustness strategies that can ensure consistent performance across a wide range of weather and lighting conditions; and the use of hardware-software co-design approaches that can maximize the computational utilization. These inventions altogether convert the theoretical abilities of perception to practical operational systems that can safely be deployed on the road. The article also covers the research directions that are emerging, such as self-supervised adaptation, fleet federation learning, and energy-conscious scheduling, which are likely to further improve the performance as well as sustainability and eventually democratize access to advanced perception-based technologies throughout the transportation sector.

Keywords: Multi-Modal Perception, Edge Computing Optimization, Autonomous Vehicles, Sensor Fusion, Environmental Robustness

1. Introduction

The autonomous vehicles industry faces unique challenges in perception system deployment that differ significantly from cloud-based machine learning implementations. A variety of sensor inputs are required by on-board perception systems. LiDAR, camera, radar, and GPS data must be processed with strict performance in latency to operate safely at highway speeds. Modern autonomous vehicle perception systems have evolved to maintain reliable environmental awareness across a wide range of road types, lighting variations, and traffic scenarios. Research in multi-modal perception integration reveals that contemporary autonomous vehicles generate substantial data volumes that must be processed with temporal consistency to maintain safe operational parameters. The heterogeneous nature of these data streams—spanning electromagnetic spectrum bands from visible light to radio frequencies—creates unique synchronization and fusion challenges that are fundamentally different from traditional cloud computing workloads, where latency requirements are measured in seconds rather than milliseconds. Recent advancements in transformer-based perception architectures have shown promise in handling these multimodal inputs, but their computational demands remain substantial when deployed in production environments [1]. Comparable challenges are reported in large-scale industrial deployments

10.48047/jocaaa.2025.34.11.37

such as NVIDIA Drive AGX and Tesla's Full Self-Driving (FSD) vision stack [13, 15], both of which illustrate the tension between perception accuracy and embedded-hardware constraints.

This real-time processing should be realized in conditions of restricted power resources and changing environmental conditions. The automotive operating environment is characterized by highly restrictive requirements of compute resources, and the perception system is usually subject to thermal design power (TDP) specifications that must deal with large temperature variations encountered during normal operation. Autonomous vehicle edge computing systems should be able to sustain constant functionality in a variety of weather parameters, such as precipitation, fog, and direct sunlight- each of which creates new perception parameters that influence the quality of sensor information. On heavy rainy mornings, radar systems become noisier, and camera-based perception has issues with dynamic range as the sun comes out of a shadow, like the entrance to a tunnel. LiDAR point clouds suffer loss in density and maximum range in poor weather conditions, which require adaptive strategies of fusion that are capable of ensuring perception reliability at different levels of sensor confidence. Field testing across diverse environmental conditions demonstrates that perception systems must dynamically adapt to these changing conditions while maintaining consistent inference timing guarantees, a requirement that substantially increases system complexity compared to controlled data center deployments [2].

Early autonomous vehicle development prioritized perception accuracy through computationally intensive neural networks. Initial deep learning architectures focused primarily on achieving benchmark superiority, often utilizing ensemble approaches and high-parameter models that delivered impressive results on standardized datasets. These models typically employed deep convolutional neural networks with specialized detection heads optimized for different perception tasks—semantic segmentation, object detection, depth estimation, and motion prediction. While these approaches achieved impressive mAP (mean Average Precision) scores on validation datasets, they generally assumed unlimited computational resources. The gap between research implementations and production requirements became apparent as these architectures were deployed to resource-constrained embedded platforms, where inference times regularly exceeded safety thresholds for highway-speed operation. The computational complexity of these early networks, particularly in multi-frame temporal fusion operations, proved challenging for automotive-grade hardware despite their theoretical accuracy advantages. This performance disparity highlighted the need for perception architectures specifically designed for edge deployment rather than merely adapted from cloud-optimized implementations [1].

This incompatibility catalyzed research into efficient deployment strategies that could preserve perception quality while meeting edge computing constraints. Changes to model architecture constituted new blocks of efficiency-oriented building blocks, including depthwise separable convolutions, attention-based pruning, and hardware-conscious neural architecture search. These methods radically rethought the trade-off between the computational complexity and the perceived accuracy, frequently being able to match the performance of larger networks with only a fraction of the floating point operations. Edge-specialized quantization techniques further reduced memory requirements through mixed-precision implementations that strategically preserved full precision only for critical network parameters. The temporal nature of perception tasks enabled innovations in recurrent computation patterns, where certain feature extraction operations could be amortized across multiple frames rather than redundantly recomputed. Critically, these optimizations were developed with production hardware constraints as primary design parameters rather than afterthoughts, leading to architectures inherently suited for automotive deployment. The field has subsequently witnessed remarkable advancements in deploying sophisticated AI pipelines on

10.48047/jocaaa.2025.34.11.37

automotive-grade GPUs and specialized inference accelerators, enabling complex perception capabilities within the resource constraints of production vehicles [2].

2. System Architecture Design

The foundation of effective edge deployment begins with modular pipeline architecture. Modern autonomous vehicle perception environments divide computational processes into specialized units that operate sensor information in sequential steps but have parallel execution paths. These architectures deploy special preprocessing threads to each sensor modality to enable the camera, LiDAR, and radar inputs to be processed simultaneously without introducing execution bottlenecks. Each of the modalities is extracted into a semantic representation, and then multi-modal fusion layers are used to combine complementary information between sensors. This modular design philosophy enables perception engineers to optimize individual components independently while preserving system-level determinism through well-defined tensor interfaces between modules. Remote sensing research demonstrates that modular architectures enable more effective integration of multi-scale feature extraction, particularly when combining aerial and ground-based perception data, a methodology that transfers effectively to autonomous vehicle deployment scenarios [3].

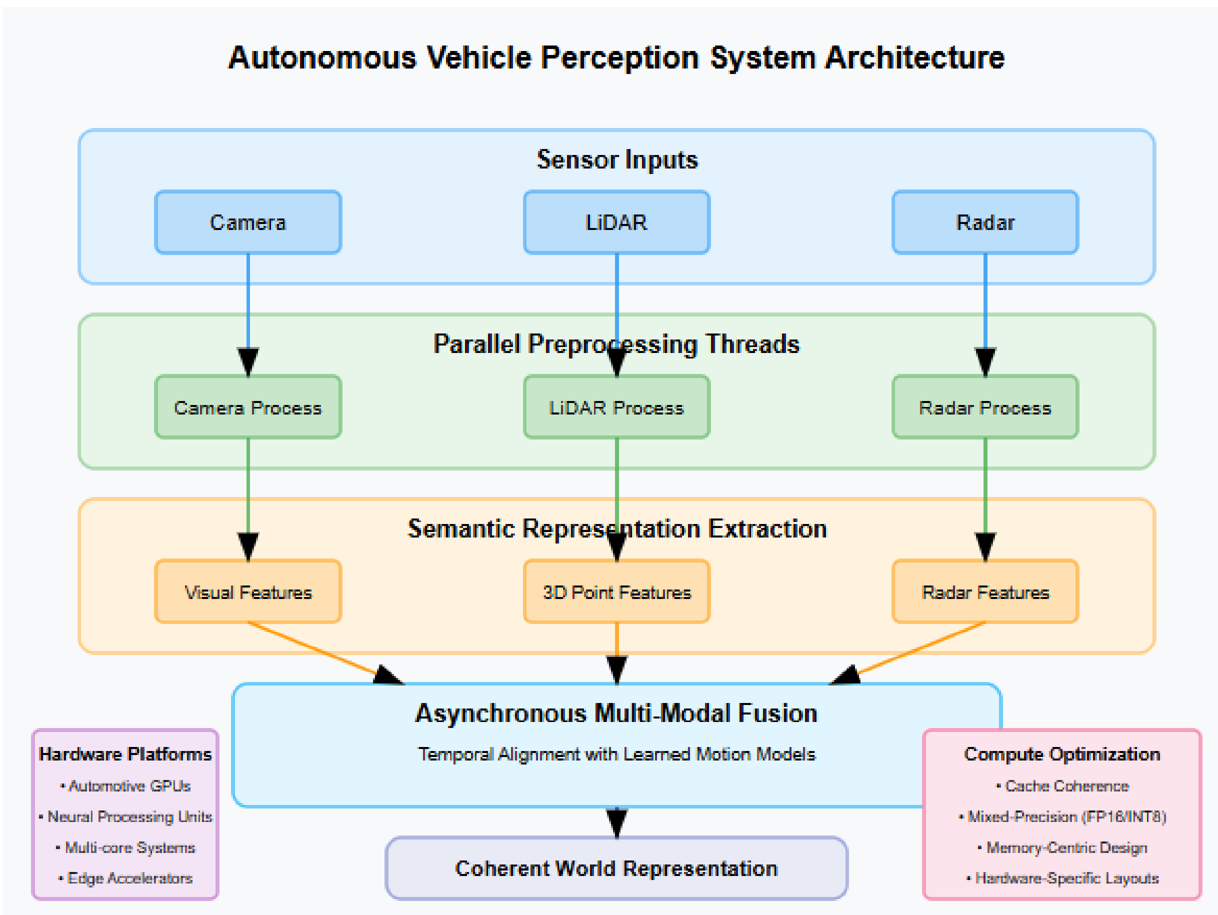

Fig. 1. Modular Pipeline Architecture for Autonomous Vehicle Perception Systems [3, 4]

10.48047/jocaaa.2025.34.11.37

Asynchronous multi-stream inference represents a significant architectural advancement, allowing different sensor modalities to process concurrently and synchronize only when fusion becomes necessary. This approach liberates high-frequency sensors from waiting for slower modalities, enabling camera streams to process at their native frame rates while LiDAR data, which typically arrives less frequently, joins the pipeline when available. Modern perception systems implement sophisticated temporal alignment mechanisms that maintain coherent world representations despite varying sensor refresh rates. These temporal fusion approaches rely on learned motion models that can predict intermediate states between sensor readings, ensuring that fusion operations combine temporally consistent information despite hardware-imposed timing differences between modalities [3]. This modular multi-stream philosophy echoes the architecture of NVIDIA DriveWorks [13] and Waymo's transformer-based perception system [14], both of which implement hardware-synchronized pipelines optimized for edge deployment.

Engineers strive to optimize compute locality to reduce the latency bottlenecks in the perception pipeline. Systems capable of enforcing intermediate tensors on-equipment, and lacking the needless CPU-GPU memory transfers, can accomplish sub-50 milliseconds perception latency, a threshold believed to make safe operation at highway velocities. Recent embedded systems architectures of autonomous vehicles have sophisticated memory management policies that emphasize cache coherence across multi-core systems. Mixed-precision computation with the use of FP16 or INT8 formats helps to minimize memory consumption, although at the cost of detection accuracy by employing extremely careful calibration schemes. Research in real-time robotics perception demonstrates that memory-centric design approaches yield substantial throughput improvements, particularly when implemented alongside hardware-specific tensor layout optimizations that maximize cache utilization on automotive-grade GPUs and neural processing units [4].

Architecture Component	Memory Efficiency	Processing Parallelism	Environmental Adaptability
Traditional Sequential Processing	Low	Low	Low
Modular Pipeline Architecture	Medium	Medium	Medium
Asynchronous Multi-Stream Inference	Medium	High	Medium
Compute Locality Optimization	High	Medium	Medium
Mixed-Precision Computation (FP16)	High	Medium	Medium
Mixed-Precision Computation (INT8)	Very High	High	Medium
Full Hardware-Specific Optimization	Very High	Very High	High

Table 1. Edge Computing Architecture: Performance Metrics for Autonomous Perception [3, 4]

3. Model Optimization Strategies

Several key optimization techniques have emerged as essential for edge deployment of perception models in autonomous vehicle systems. Structured pruning has revolutionized model compression by targeting architectural elements rather than individual weights. Unlike random weight elimination that creates

10.48047/jocaaa.2025.34.11.37

irregular sparsity patterns, the structured pruning methodically removes entire convolutional channels based on contribution metrics derived from activation statistics and gradient magnitudes. This approach preserves computational density that aligns with modern tensor processing hardware, enabling efficient execution on automotive-grade accelerators. The channel-wise nature of structured pruning maintains memory access coherence critical for embedded systems with limited cache hierarchies. Research on autonomous vehicle perception models indicates that carefully implemented structured pruning can reduce computational requirements by 50-65% while maintaining object detection and segmentation performance within acceptable safety margins. The most effective implementations utilize iterative pruning schedules that gradually remove channels while allowing remaining parameters to compensate through intermittent fine-tuning phases, preserving the most information-rich feature extractors throughout the network architecture [5].

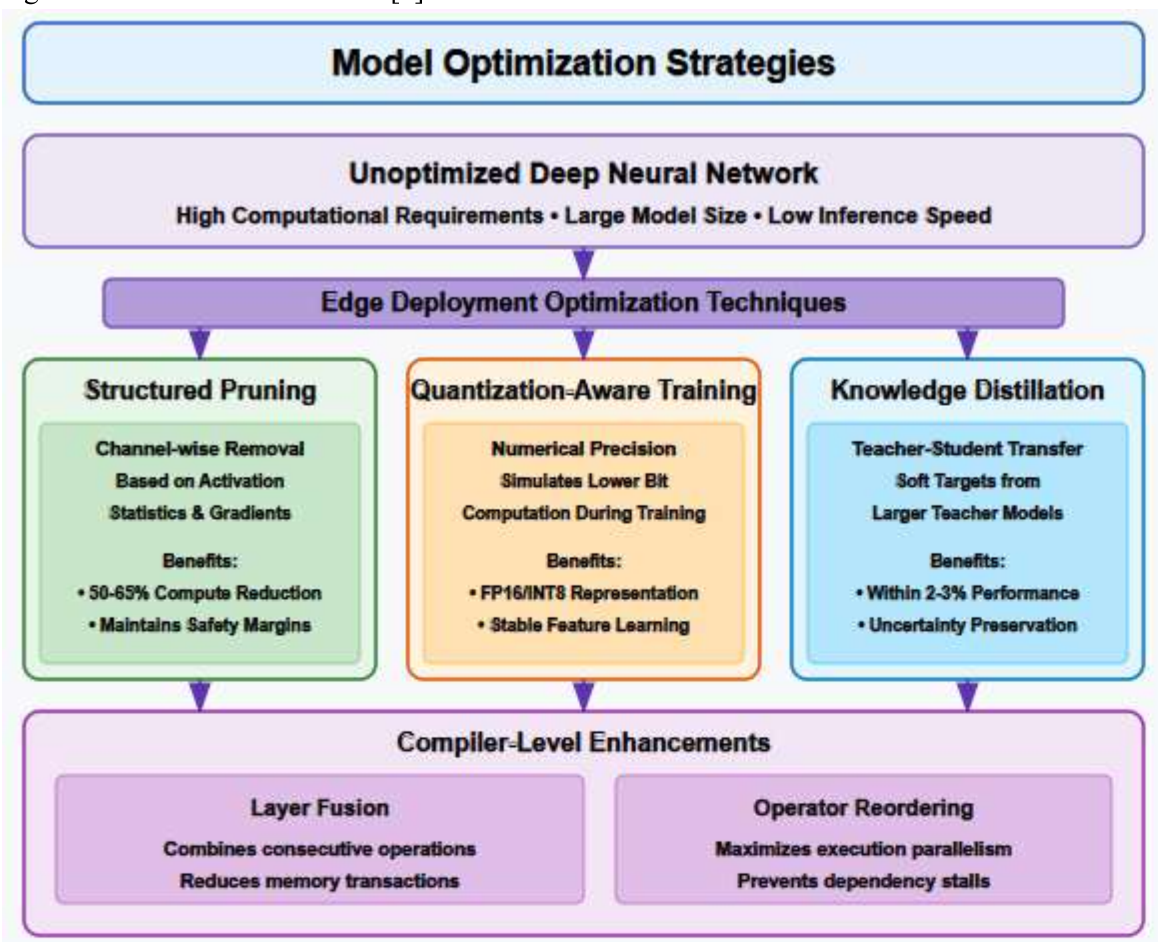

Fig. 2. Model Optimization Strategies for Edge Deployment in Autonomous Vehicles [5, 6]

Quantization-aware training represents another crucial optimization that addresses the numerical precision dimension of model efficiency. By incorporating simulated quantization effects during the training process, networks learn robust feature representations that remain stable despite reduced precision. This approach extends beyond post-training quantization by allowing the network to adapt its parameter distributions specifically for lower-bit computation. Modern implementation frameworks integrate quantization nodes directly into the computational graph, simulating integer operations in the forward pass while maintaining higher precision for gradient computation. Knowledge distillation complements

10.48047/jocaaa.2025.34.11.37

these techniques by transferring learned representations from high-capacity "teacher" networks to compact "student" models. The distillation process typically employs soft targets derived from teacher logits, enabling the student to learn both classification boundaries and confidence distributions. This approach proves particularly valuable for safety-critical perception tasks where uncertainty quantification remains as important as point predictions. Research demonstrates that properly distilled models can achieve detection performance within 2-3% of their full-sized counterparts while requiring significantly fewer parameters and computational operations [6].

These optimization techniques, when combined with compiler-level enhancements such as layer fusion and strategic operator reordering, form the foundation of edge-efficient perception for autonomous vehicle systems. Layer fusion eliminates redundant memory operations by combining consecutive network operations into optimized kernel implementations. Operation rearrangement is an algorithm to rearrange the computational graphs to achieve the maximum execution parallelism on the multi-core architecture without introducing dependency-related stalls. These methods combined can permit the execution of complicated perception models operating within the harsh latency limits needed to operate safely on the road.

Optimization Technique	Computation Reduction	Model Size Reduction	Accuracy Preservation	Implementation Complexity	Memory Bandwidth Savings
Unoptimized Model	None	None	Very High	Very Low	None
Random Weight Pruning	Low	Medium	Medium	Medium	Very Low
Structured Pruning	High	High	High	High	Medium
Post-Training Quantization	Medium	High	Medium	Low	High
Quantization-Aware Training	High	Very High	High	Medium	Very High
Knowledge Distillation	Very High	Very High	High	High	High
Layer Fusion	Medium	Very Low	Very High	Medium	High
Operator Reordering	Low	None	Very High	Medium	Low
Combined Optimization Stack	Very High	Very High	High	Very High	Very High

Table 2. Edge AI Optimization: Qualitative Comparison for Autonomous Perception [5, 6]

4. Sensor Fusion and Data Management

Multi-modal perception for autonomous vehicle systems faces substantial data throughput challenges that must be addressed through strategic management approaches. Contemporary sensor arrays generate massive data volumes that exceed the processing capabilities of edge computing platforms when handled

10.48047/jocaaa.2025.34.11.37

naively. LiDAR sensors in modern days with 128 or 256 scan channels produce dense point clouds with millions of spatial positions per frame and coded with position, intensity, and time data. At the same time, multiple gigabits per second of uncoded visual information are generated by high-definition camera arrays with resolutions of over 8 megapixels per sensor. This combined sensor throughput creates potential bandwidth bottlenecks between data acquisition systems and processing hardware. Edge computing platforms address these challenges through sophisticated data reduction techniques implemented at multiple pipeline stages. Region-of-interest filtering employs adaptive attention mechanisms that prioritize processing resources toward areas with high information content—typically regions containing dynamic objects, road boundaries, or traffic infrastructure. This filtering occurs through both hardware-accelerated preprocessing and neural attention mechanisms within perception networks. Adaptive frame subsampling further reduces computational load by dynamically adjusting sensor refresh rates based on scene complexity and vehicle dynamics, decreasing processing frequency during relatively static scenarios while maintaining high temporal resolution when navigating complex traffic situations. These approaches collectively enable efficient resource allocation while preserving perception quality in safety-critical regions of the driving environment [7].

Advanced asynchronous fusion frameworks represent a complementary innovation that addresses the inherent timing mismatches between heterogeneous sensor modalities. Rather than enforcing rigid synchronization requirements that introduce artificial processing delays, modern architectures implement flexible temporal alignment strategies that maximize information utilization while maintaining algorithmic determinism. These frameworks construct persistent spatio-temporal feature maps that accumulate sensory information with appropriate timestamp annotations, enabling fusion algorithms to incorporate measurements despite variable arrival patterns. Motion compensation modules use acquired kinematic models to combine past sensor readings into the present coordinate reference frames, including ego-vehicle movement and anticipated object motion. This dynamic alignment method allows the perception system to have consistent world representations even when sensor refresh rates vary drastically: cameras at 30-60 Hz, LiDAR at 10-20 Hz, and radar at 13-40 Hz. The quality of temporal alignment has a direct effect on detection robustness, especially when dealing with high-speed objects, because even small synchronization errors can lead to significant position differences. Experimental validation across diverse driving conditions demonstrates that sophisticated temporal fusion strategies significantly outperform rigid synchronization approaches in both detection accuracy and computational efficiency. This capability proves particularly valuable for autonomous vehicles including commercial transportation, where consistent perception performance must be maintained across variable traffic conditions, highway speeds, and extended operational durations [8]. Comparable asynchronous-fusion strategies are implemented in NVIDIA Drive Perception SDK [13] and Waymo's temporal alignment modules [14], confirming a broader industry trend toward learned motion-compensated fusion across heterogeneous refresh rates.

Component	Data Volume	Refresh Rate	Processing Complexity	Environmental Resilience	Spatial Resolution	Temporal Precision
Camera Sensors	Very High	High	Medium	Low	High	Medium
LiDAR Sensors	High	Medium	High	Medium	Very High	Low
Radar Sensors	Low	Medium	Low	High	Low	High

10.48047/jocaaa.2025.34.11.37

Adaptive Frame Subsampling	Low	Variable	Low	Medium	Medium	Variable
Rigid Synchronization	Medium	Low	Medium	Low	High	Low
Asynchronous Fusion	Low	High	High	High	High	High
Temporal Alignment with Motion Compensation	Low	High	Very High	Very High	Very High	Very High

Table 3. Multi-Modal Sensor Integration: Qualitative Performance Matrix for Autonomous Vehicles [7, 8]

5. Environmental Robustness

The conditions to which perception systems are exposed on the road are truly extraordinary, putting the hardware, software, and algorithms to the test. Autonomous vehicles operating across diverse environments routinely experience a wide range of conditions, such as dust storms, which block optical sensors, rain conditions (light to heavy), which weaken LiDAR signals, direct sunlight, which causes the camera to be overexposed, and the presence of mechanical vibration, which destabilizes sensor calibration. These environmental influences lead to progressive and abrupt deterioration in sensor data quality, which may compromise the reliability of perception if not carefully managed. The efforts of the state-of-the-art architectures apply to these issues with self-calibration perception modules that continuously check sensor performance indicators and dynamically adapt fusion network weights depending on the perceived reliability. These adaptive systems implement sophisticated sensor quality assessment algorithms that detect specific degradation patterns—LiDAR point cloud density reduction, camera image blur, radar signal attenuation—and modify perception pipeline parameters accordingly. For example, when fog reduces LiDAR intensity and range, the system compensates by increasing reliance on camera-based depth inference and radar returns for distant object detection. Similarly, during camera overexposure events from direct sunlight, perception systems increase dependence on LiDAR-derived geometric features while reducing the weight assigned to color and texture information. This dynamic adaptation ensures consistent perception performance across varying environmental conditions without requiring separate models for each scenario, significantly increasing system robustness for extended autonomous operations [9].

Domain-adaptive training approaches also augment environmental resilience with systematic exposure to different operating environments in training the models. Conventional supervised methods of learning tend to have a serious decline in performance in situations that are not consistent with training distributions, which is especially difficult to address in autonomous vehicle deployment, which faces the need to operate on geographic scales, weather conditions, and infrastructure diversity. Sophisticated training systems overcome this issue by modeling the various lighting, weather, and environmental conditions with complex render engines and sensor models based on physics. Some of the techniques used include style transfer augmentation, which allows the generation of training examples that preserve semantic information but differ in visual appearance features like illumination, contrast, and texture. Domain randomisation is an intentional source of variability in training data, which compels perception models to concentrate on the properties of an object but not on the patterns of a dataset. Progressive

10.48047/jocaaa.2025.34.11.37

domain adaptation frameworks gradually bridge synthetic and real-world distributions through curriculum learning approaches, where models initially train on clearly separated domains before encountering increasingly challenging transfer scenarios. These methodologies collectively reduce the generalization gap between controlled development environments and unpredictable operational conditions. Field deployment data demonstrates that domain-adaptive training yields measurable safety improvements, including up to 45% reduction in false-negative rates during adverse weather and more consistent object detection confidence scores across environmental variations. These adaptations transform perception systems from brittle, condition-specific implementations into robust modules capable of reliable operation across the full spectrum of environmental challenges encountered in autonomous driving, with particular benefits for commercial vehicles like autonomous trucks that must operate reliably in all conditions [10].

Environmental Challenge	Camera Impact	LiDAR Impact	Radar Impact	Overall Perception Risk
Dust/Particulates	Very High	High	Low	High
Light Rain	Medium	Medium	Very Low	Medium
Heavy Rain	High	Very High	Low	Very High
Snow	Very High	Very High	Medium	Very High
Direct Sunlight	Very High	Low	Very Low	High
Low Light/Night	Very High	None	None	High
Fog	High	Very High	Low	Very High
Mechanical Vibration	Medium	High	Low	Medium
Temperature Extremes	Low	Medium	Medium	Medium

Table 4. Environmental Challenges Impact Matrix [9, 10]

6. Future Research Directions

The field of autonomous vehicle perception continues to advance toward self-supervised adaptation and federated learning approaches, which represent promising frontiers for enhancing edge-deployed intelligence. Self-supervised adaptation enables perception systems to continuously refine their models based on operational experience without requiring explicit human labeling. These approaches leverage inherent structure in sensor data to generate training signals, such as using temporal consistency between frames to validate detection quality or employing cross-modal verification where one sensor modality provides supervision for another. Federated learning methodologies extend this paradigm to fleet-scale improvement by enabling distributed model enhancement across multiple vehicles while maintaining data privacy. Rather than centralizing sensitive operational data, federated approaches distribute training computations across edge devices, sharing only model updates rather than raw sensor recordings. This architecture preserves privacy while allowing the collective fleet intelligence to rapidly improve rare-event detection capabilities such as unusual road hazards, uncommon traffic scenarios, or edge-case weather conditions. Preliminary implementations demonstrate that federated perception learning can accelerate model improvement for rare categories by 3-5× compared to centralized approaches while maintaining strict data locality requirements. These advances particularly benefit autonomous vehicle fleets, especially in commercial transportation where extensive highway miles provide abundant learning

10.48047/jocaaa.2025.34.11.37

opportunities for continuous improvement. Future research will focus on communication-efficient update protocols, differential privacy guarantees, and heterogeneous hardware accommodations to make federated learning practical across diverse vehicle fleets with varying computational capabilities [11]. Industrial deployments such as Waymo's fleet-scale learning platform [14] are early demonstrations of these principles, underscoring their practical feasibility for commercial fleets.

Another important research direction is energy-conscious scheduling systems, which will improve the sustainability of the environment as well as operational efficiency. The architectures used today generally operate at fixed computational intensity independent of the context under which they are being driven, and consume as much power as possible, even in cases where a simplified processing model would be appropriate. Further studies are done on a context-adaptive computation, which changes the model complexity and sensor usage dynamically according to the environmental conditions, velocity, traffic, and routing data. On highways that have low traffic and good weather conditions, simplified perception models and lower sensor rates can be activated with high savings in energy. Conversely, complex urban intersections or adverse weather conditions would trigger full-capacity perception with maximum sensor utilization. Early prototype results show a reduction in energy consumption of 30-45 percent during normal highway operations with no reduction in safety margins. In the broader perspective of impact, dependable real-time perception systems for autonomous vehicles will result in the reduction of fatigue-related accidents, improved fuel efficiency due to efficient route planning, and enhanced supply chain resilience. These benefits are particularly significant for heavy-duty vehicles like autonomous trucks, where efficiency gains have outsized economic and environmental impacts. As edge-optimized technologies become more accessible, the safety features of autonomous systems can be democratized and delivered to smaller logistics companies at no cost. These accessibility enhancements not only advance industry application of perception-enhanced safety systems faster than full autonomy can, but they also generate immediate benefits in society by reducing accidents and improving the efficiency of the current fleet operated by humans. The convergence of these technologies—self-supervised learning, federated improvement, and energy-aware computation—establishes a foundation for sustainable and continuously improving perception systems that can adapt to changing environments while maintaining strict operational safety requirements [12].

Conclusion

The process of enhancing autonomous vehicle perception systems and developing production-ready applications out of laboratory prototypes is a stunning collaboration of algorithmic design and engineering realism. Researchers and engineers have successfully addressed the underlying problem of the deployment of complex neural networks to the edge computing platform through careful architectural design, advanced model optimization, adaptive sensor fusion, and environmental robustness strategies. Such developments allow us to have dependable perception in real-world scenarios as well as observe the rigid latency demands required in diverse driving environments. The adaptive, modular character of modern perception systems gives the opportunity to continuously enhance the system based on operational experience, establishing the basis for increasingly robust autonomy. In addition to the technical success, these systems testify to the fact that artificial intelligence can be introduced into safety-critical infrastructure in a responsible manner by taking into account operational needs and environmental conditions. These technologies will have a massive impact on society through reduced accident rates, increased efficiencies in logistics, and improved supply chain resilience, particularly in commercial

10.48047/jocaaa.2025.34.11.37

transportation with autonomous trucks, as they mature and become more accessible. The field also shows that computational constraints can be used to generate innovation, instead of constraining it, leading to perception systems being designed not only to perform well on benchmarks but to have long-term operational confidence in dynamic real-world conditions.

References

[1] Jiaqi Peng et al., "Towards Latency-aware 3D Streaming Perception for Autonomous Driving," arXiv:2504.19115v1, 2025. [Online]. Available: https://www.arxiv.org/html/2504.19115v1

[2] Nima Ghafoorianfar and Mehdi Roopaei, "Environmental Perception in Autonomous Vehicles Using Edge Level Situational Awareness," ResearchGate, 2020. [Online]. Available: https://www.researchgate.net/publication/339903761_Environmental_Perception_in_Autonomous_Vehicles_Using_Edge_Level_Situational_Awareness

[3] Yuxiao Zhang et al., "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," ISPRS Journal of Photogrammetry and Remote Sensing, Volume 196, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271622003367

[4] Kamilya Smagulova et al., "Efficient and real-time perception: a survey on end-to-end event-based object detection in autonomous driving," Frontiers in Robotics and AI, 2025. [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1674421/full

[5] Byeongseon Park et al., "Progressive Pruning of Light Dehaze Networks for Static Scenes," Applied Sciences, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/23/10820

[6] Wolfgang Roth et al., "Resource-Efficient Neural Networks for Embedded Systems," Journal of Machine Learning Research, 25, 2024. [Online]. Available: https://www.jmlr.org/papers/volume25/18-566/18-566.pdf

[7] Sihai Tang et al., "Vehicular Edge Computing for Multi-Vehicle Perception," 2021 Fourth International Conference on Connected and Autonomous Driving (MetroCAD), 2021. [Online]. Available: https://www.ieeexplore.ieee.org/document/9499333

[8] Hui Qian et al., "A Review of Multi-Sensor Fusion in Autonomous Driving," Sensors, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/19/6033

[9] Ziqiang Zheng et al., "Robust Perception Under Adverse Conditions for Autonomous Driving Based on Data Augmentation," ResearchGate, 2023. [Online]. Available: https://www.researchgate.net/publication/374821822_Robust_Perception_Under_Adverse_Conditions_for_Autonomous_Driving_Based_on_Data_Augmentation

[10] Bamidele Matthew et al., "Domain Adaptation Techniques for Transferring Perception Models from Simulation to Real-World Road Environments," ResearchGate, 2025. [Online]. Available: https://www.researchgate.net/publication/395234289_Domain_Adaptation_Techniques_for_Transferring_Perception_Models_from_Simulation_to_Real-World_Road_Environments

[11] Mohab M. Eid Kishawy et al., "Federated learning system on autonomous vehicles for lane segmentation," Scientific Reports, Volume 14, Article Number 25029, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-71187-8

[12] Verónica Bolón-Canedo et al., "A review of green artificial intelligence: Towards a more sustainable future," Neurocomputing, Volume 599, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224008671

[13] NVIDIA Corporation, NVIDIA DriveWorks SDK 5.0: Building Perception and Sensor Fusion for Autonomous Driving, Whitepaper, 2024. [Online]. Available: https://www.developer.nvidia.com/driveworks

10.48047/jocaaa.2025.34.11.37

[14] W. Maddern et al., "End-to-End Transformer-Based Multi-Sensor Fusion for Autonomous Vehicles," Waymo Research, 2023. [Online]. Available: https://www.waymo.com/research
[15] A. Karpathy, "Tesla FSD Architecture – Vision, Planning, and Hardware Co-Design," Tesla AI Day Whitepaper, 2022. [Online]. Available: https://www.tesla.com/AI