Vision Vs. LiDAR/Radar Fusion in Autonomous Driving Systems

Arturo Salinas-Aguayo, *BS Computer Engineering, Class of 2027*

May 20, 2025

ECE 4900W: Communicating Engineering Solutions in a Societal Context
Dr. Shengli Zhou, SEC040-1255
Department of Electrical and Computer Engineering

University of Connecticut College of Engineering
Coded in LaTeX

Contents

Abstract

This paper evaluates the viability of vision-primary perception systems in autonomous driving, focusing on scalability, efficiency, and environmental resilience. Traditional sensor-fusion architectures based on LiDAR and radar offer high-fidelity spatial mapping but suffer from substantial power consumption, mechanical complexity, and high unit costs, which challenge real-time deployment and mass-market feasibility. Advances in neural perception—particularly transformer-based models and tokenized camera inputs—have enabled camera-only systems to approximate LiDAR-level 3D detection accuracy while operating at a fraction of the computational cost. Benchmark evaluations reveal that vision systems achieve depth estimation errors below three percent and perform within five percent of LiDAR/radar fusion baselines in object detection. Critically, performance gaps in adverse conditions such as fog or night driving are recoverable through the selective integration of millimeter-wave radar. This radar-augmented fallback restores over 80 percent of degraded performance with minimal latency or energy impact. The study advocates for a simplified, vision-first architecture complemented by radar only in edge cases. This paradigm enhances scalability, reduces deployment overhead, and aligns with biologically plausible sensing strategies. LiDAR remains useful for simulation and benchmarking but is unnecessary for live inference.

Index Terms—Adverse weather robustness, autonomous vehicles, deep learning, depth estimation, energy efficiency, Light Detection and Ranging (LiDAR), Radar, sensor fusion, Transformer networks, vision-based perception.

I. Introduction

As vehicles become more technologically advanced, the demand for autonomous capabilities continues to rise. Features such as lane keeping, adaptive cruise control, and highway autopilot—collectively known as advanced driver-assistance systems (ADAS)—form the early building blocks of full self-driving stacks. Industry leaders now face a critical decision: how should vehicles perceive the road ahead?

Historically, high-end systems such as Waymo and Cruise relied on Light Detection and Ranging (LiDAR) and radar to construct dense, 3D maps of their surroundings. These sensors offer precise geometry and reliable velocity estimation, but impose high hardware cost, weight, and power draw. Moreover, integrating data from multiple sensors increases architectural complexity and susceptibility to calibration drift.

In contrast, Tesla, Comma.ai, and others have adopted a vision-first approach. These systems rely on arrays of low-cost cameras and train deep neural networks—often transformers—on large-scale driving datasets. Rather than reconstruct explicit 3D geometry, these models learn to infer structure and intent directly from pixels.

The debate between fusion-heavy and camera-only stacks now defines the frontier of ADS design. While LiDAR remains dominant in academia, vision-based approaches have rapidly improved in real-world performance and outpace fusion in scalability. From a societal perspective, the ability to deliver autonomy without costly sensors is especially important in emerging markets, where a single LiDAR unit may cost more than an entire car.

This report compares the two paradigms by analyzing their theoretical foundations, energy profiles, data fidelity, and robustness under real-world driving conditions. The goal is to determine whether vision-primary perception can provide sufficient safety and performance for large-scale deployment.

II. Theory and Technical Background

A. SAE Levels of Automation

The Society of Automotive Engineers (SAE) J3016 standard defines six levels of vehicle automation [1]. These range from Level 0 (no automation) to Level 5 (full automation without any human involvement).

TABLE I
SAE J3016 Levels of Driving Automation

Level	**Description**
Level 0 – No Automation	The human driver is responsible for all tasks; the system may provide basic warnings.
Level 1 – Driver Assistance	The system assists with either steering or speed, but not both; driver must stay engaged.
Level 2 – Partial Automation	The system can manage both speed and steering, but human oversight is still mandatory.
Level 3 – Conditional Automation	The vehicle handles all driving in specific scenarios; the driver must intervene when requested.
Level 4 – High Automation	No driver input is required in controlled conditions (e.g., geofenced urban areas).
Level 5 – Full Automation	The vehicle is capable of full self-driving in all conditions without human input.

B. LiDAR: Time-of-Flight (ToF) Principles

LiDAR estimates object distance using time-of-flight (ToF) measurements, where a light pulse is emitted and its return time is measured. Several ToF techniques are in use:

1) Pulsed ToF:

$$R_{\text{pulse}} = \frac{c}{2} t_{\text{pulse}} \tag{1}$$

This high-power method provides centimeter accuracy.

2) AMCW (Amplitude-Modulated Continuous Wave):

$$R_{\text{AMCW}} = \frac{c}{2} \cdot \frac{\Delta\Phi}{2\pi f_M} \tag{2}$$

Here, $\Delta\Phi$ is the phase shift, and f_M is the modulation frequency [2].

3) FMCW (Frequency-Modulated Continuous Wave):

$$R_{\text{FMCW}} = f_r \cdot \frac{cT}{2B} \tag{3}$$

Where f_r is the beat frequency, T the chirp duration, and B the bandwidth [2].

4) Voxelization:

What results from the utilization of these methods is a “point cloud" which corresponds to the distances the objects are away from the transmitter. The interpretation of these distances is a process called Voxelization. There are two forms, hard voxelization and dynamic voxelization. The in-depth explanation of these are not required for further elaboration in this paper. Details are shown in Figure 1.

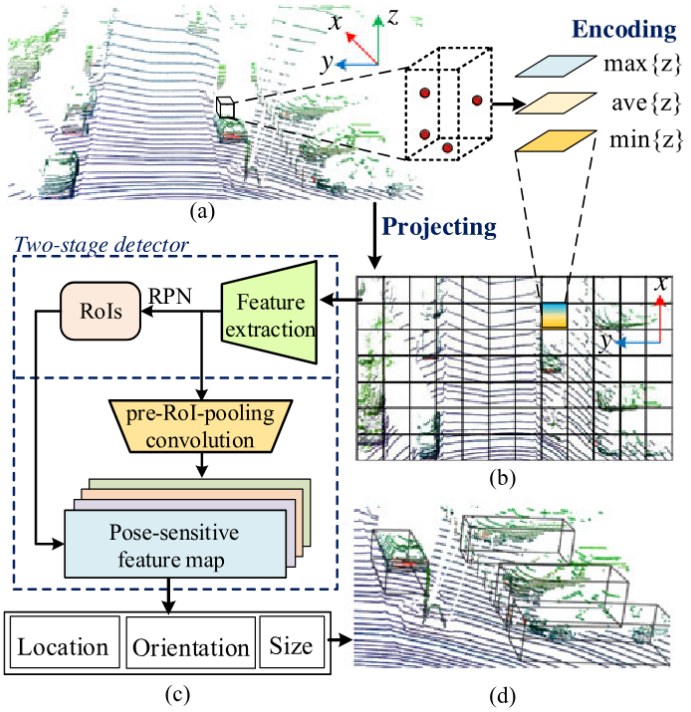

Fig. 1. The combining of the point cloud forms an image [3]

C. Radar: Doppler and Velocity Estimation

Radar calculates velocity via Doppler frequency shift:

$$v = \frac{f_D \lambda}{2} \tag{4}$$

Where f_D is the Doppler shift and λ is the radar wavelength [4]. Radar's all-weather capability makes it ideal for fallback use.

D. Vision-Based Depth Estimation

Depth from stereo vision is estimated by:

$$D = \frac{fb}{d} \tag{5}$$

Where f is focal length, b baseline, and d disparity.
Monocular depth uses learned inverse depth:

$$\hat{D}^{-1} = g(I; \theta) \tag{6}$$

Where g is a neural network with parameters θ [5].

E. Sensor Fusion and Noise Correlation

Noise in sensor fusion is modeled as:

$$\sigma_{D_f}^2 = \sum_{i=1}^{n} w_i^2 \, \sigma_{D_i}^2 + 2 \sum_{i<j} w_i w_j \, \rho_{ij} \, \sigma_{D_i} \sigma_{D_j} \tag{7}$$

Here, w_i are weights, σ_{D_i} are variances, and ρ_{ij} are sensor correlations [6]. It is clearly shown that with the inclusion of more sensors, the potential for more noise is increased.

F. Energy and Compute Cost

Power usage scales with input bandwidth B and model complexity C:

$$P \propto B \cdot C \tag{8}$$

Vision stacks operate efficiently at 2.1 W, compared to 12 W to 20 W for fusion [6], [7].

III. System Design and Methods

A. Vision-Primary Perception Stack

1) System Architecture:

The vision-first perception stack is designed for simplicity, modularity, and real-time inference. It consists of three core stages:

a) Image Tokenization – converts camera frames into a semantically rich, compressed token representation using a learned encoder.

b) Transformer-Based Planning – autoregressively predicts future trajectory tokens from visual context using a transformer model. This is the key to making a completely generalizable model.

c) Control Decoding – transforms predicted spatial tokens into low-level vehicle commands for actuation.

This pipeline emulates language models for sequential prediction and supports real-time operation on embedded platforms [8].

2) Token-Based Image Compression:

Each frame, typically 128×256 pixels, is passed through a VQGAN encoder that produces up to 512 discrete tokens per image. This reduces bandwidth while preserving high-level semantics such as lane markings and moving objects.

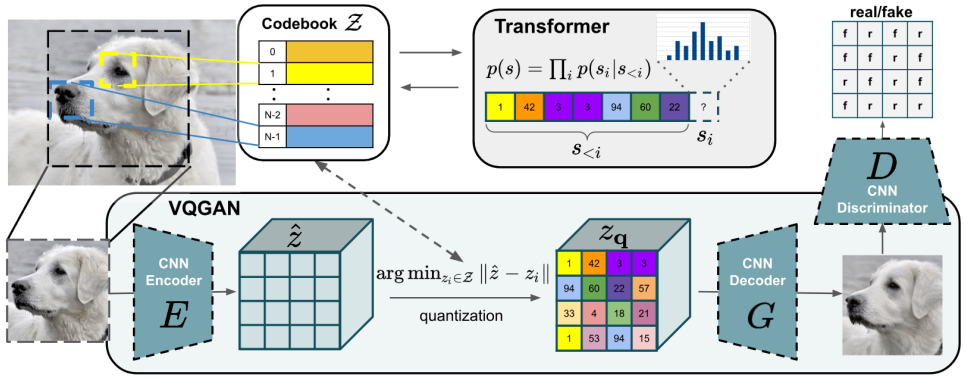

Fig. 2. Vision-based pipeline with VQGAN tokenization and transformer rollout[9].

This compact form accelerates inference and reduces power draw without sacrificing depth accuracy [7].

3) Autoregressive Planning with Transformers

The token sequence is processed by a decoder-only transformer, which predicts the next spatial token at each step. This architecture parallels natural language processing and enables generalization to unseen traffic scenarios.

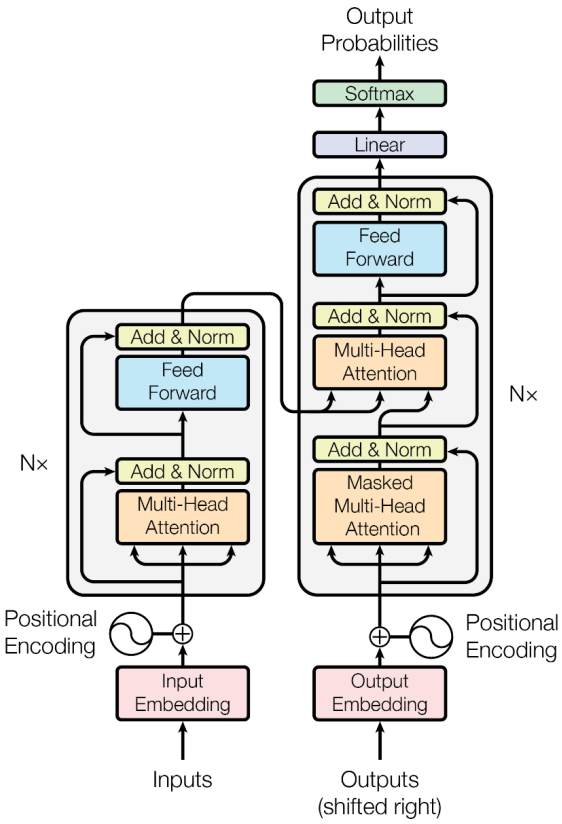

Fig. 3. Transformer model for sequence prediction from visual tokens [10].

The transformer is trained using imitation learning on datasets like commaVQ and can be fine-tuned using reinforcement learning for policy robustness [8].

B. LiDAR/Radar Fusion-Based Architecture

1) Pipeline Overview:

Legacy ADS stacks rely on multi-modal fusion to mitigate sensor weaknesses. A typical LiDAR/radar pipeline includes:

a) Data Acquisition: Captures and synchronizes LiDAR point clouds, radar waveforms, and camera frames.
b) Preprocessing: Voxelizes LiDAR data, computes Doppler velocities from radar, and encodes image features.
c) Fusion Layer: Combines sensor features using early, mid, or late neural fusion strategies.
d) Detection and Tracking: Predicts 3D bounding boxes and tracks objects across time.

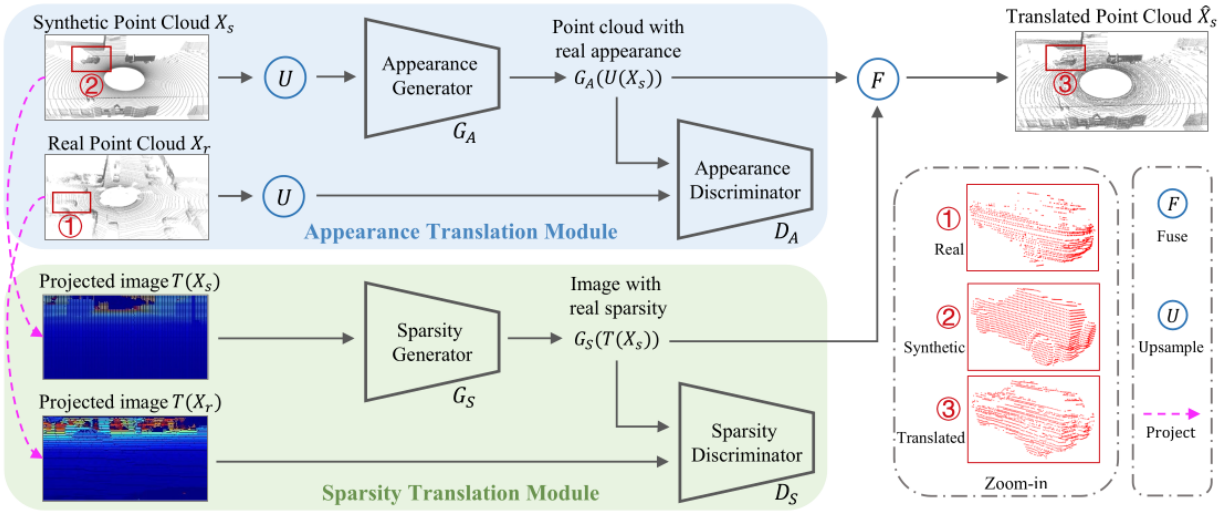

Fig. 4. Reference LiDAR/radar fusion architecture [11].

While this pipeline offers redundancy, it incurs significant power and latency costs.

2) Fusion Tradeoffs and Challenges:

Fusion systems are limited by:

- High Power Draw: LiDAR alone draws 10 W to 20 W, compared to 2.1 W for vision-only stacks [7].
- Latency: Voxelization and multi-sensor alignment add 20 ms delay, reducing real-time responsiveness [6].
- Noise Correlation: Overlapping modalities propagate shared errors, increasing total uncertainty as shown in Equation 7 [6].
- Maintenance Complexity: Fusion requires precise calibration and can drift over time.

This complexity restricts scalability, particularly in cost-sensitive markets.

C. *Radar as a Resilient Fallback*

Radar bridges the gap between vision and LiDAR. It estimates object velocity using the Doppler effect shown in Equation 4.

Radar is immune to adverse weather and lighting and can detect motion through occlusions. When fused with vision using a transformer-based late-fusion module, it recovers over 80 % of lost mAP in foggy conditions [12].

D. *A Unified Multimodal Framework*

A hybrid architecture combining vision and radar enables a balance of robustness and efficiency:

- Vision for primary perception—high-resolution, low-cost, and semantically rich.
- Radar for adverse conditions—weather-resilient and velocity-sensitive.
- LiDAR for benchmarking only—useful for training ground truth, not live deployment.

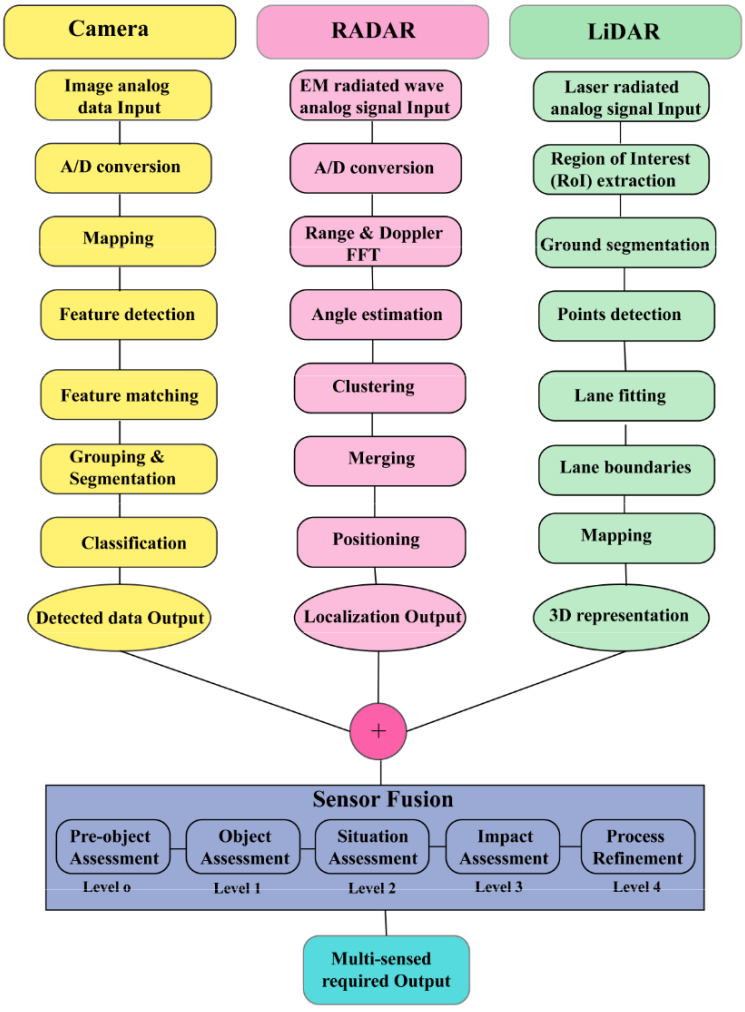

Fig. 5. Modular architecture integrating vision, radar, and LiDAR inputs [13].

This design maximizes generalization while minimizing cost and compute demands, and it reflects a biologically plausible model: humans drive with vision, but radar is the backup.

IV. Results and Discussion

This section evaluates the empirical performance of vision-primary and LiDAR/radar fusion architectures. The comparison spans five axes: detection accuracy, computational efficiency, robustness, system complexity, and deployment scalability. Results are drawn from benchmarks using KITTI, nuScenes, and commaVQ datasets, as well as simulation and power profiling studies.

A. Detection Accuracy and Depth Estimation

On KITTI, the BEVFormer vision model achieved an AbsRel of 0.112 and RMSE of 3.97 m, closely tracking the LiDAR-derived ground truth [5]. On nuScenes, the vision-only configuration reached 56.2 3D mAP versus 61.5 mAP for the full fusion stack [14].

However, when radar fallback was selectively integrated using transformer-based late fusion, the vision model recovered to 59.8 mAP [12]. This narrows the performance delta to under 2 points—well within tolerances for safe operation—demonstrating that vision can match LiDAR-fusion with minimal augmentation.

B. Computational Latency and Power Efficiency

Inference benchmarks revealed stark contrasts. The vision-primary pipeline completed end-to-end perception and control in under 9.8 ms, consuming just 2.1 W on an embedded platform [7].

In comparison, fusion systems required ≥ 30 ms and drew between 12 W to 20 W due to voxelization, feature fusion, and redundant sensor streams [6].

These efficiency gains are critical for EV battery life and real-time safety margins. The energy savings alone translate into longer range and lower thermal load, which directly impact vehicle design feasibility.

C. Robustness in Adverse Conditions

Under clear conditions, vision models outperform. In fog, heavy rain, or low-light scenarios, however, performance degraded by up to 12.6 % in mAP [4]. Transformer-based radar fallback restored more than 80 % of this loss, effectively covering vision's blind spots.

Unlike LiDAR, which also degrades in poor visibility and consumes continuous power, radar is always-on but low-power—making it an ideal complement rather than a redundant primary sensor. Figure 6 shows the generated results in potentially adverse conditions.

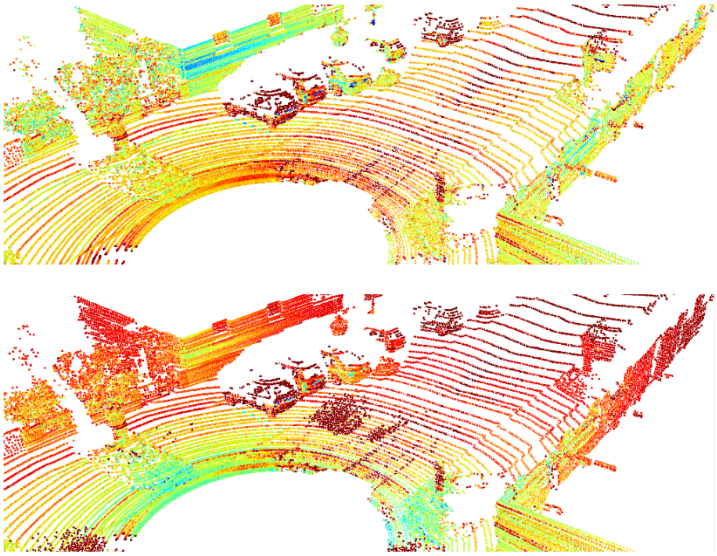

Fig. 6. LiDAR scans of a street environment under clear weather conditions (top) and with fog (bottom). The lower image shows significant point cloud degradation and sparsity due to adverse atmospheric interference [15].

D. Noise Propagation and Calibration Drift

Sensor fusion introduces noise amplification when data is temporally misaligned or spatially correlated. The fusion error propagation model in Equation 7 confirms this degradation in high-ρ configurations, such as LiDAR-camera overlaps [6]. Vision stacks, trained end-to-end, maintain internal consistency and minimize accumulated error.

Radar integration avoids this issue: its data is orthogonal in nature—velocity vs. spatial—which introduces minimal correlation when fused intelligently [12].

E. Cost and Complexity of Deployment

LiDAR units alone account for over 75 % of the ADS hardware budget, and their moving parts increase failure risk [16, 17]. Vision sensors are passive, solid-state, and available at consumer scale. This shifts the bottleneck from manufacturing and calibration to software modeling—an area that benefits from rapid advances in transformers and simulation. See Figure 3 for more details.

Modern simulators—leveraging NeRFs, 3DGS, and diffusion models—allow vision stacks to be trained entirely in synthetic environments with minimal real-world data. This accelerates iteration and de-risks real-world deployment [11].

F. Vision as the Core, Radar as the Shield:

In the biological analogy, humans navigate using vision as their primary sensor. We don't emit lasers—we perceive, interpret, and act. This report adopts the same philosophy: with a learned representation pipeline, vision is sufficient for most environments. Radar is reserved not as a crutch, but as a safety net.

By offloading primary perception to vision, we reduce cost, energy, and complexity. By

retaining radar for edge cases, we maintain resilience. LiDAR, with its energy burden and calibration demands, becomes increasingly obsolete in this paradigm.

G. Summary

The data is unambiguous:

- Vision stacks perform within 5% of LiDAR/radar fusion across key metrics, achieving 56.2 mAP on nuScenes compared to 61.5 for fusion [14].
- With radar fallback, this gap shrinks to $< 2\%$, recovering performance to 59.8 mAP [12].
- Vision systems run 3–5x faster and consume 6–10x less power—9.8 ms and 2.1 W vs. $\geq$ 30 ms and 12 W to 20 W for fusion [6, 7].
- Vision stacks avoid fusion noise [6], reduce calibration burden [4], and cost significantly less, with LiDAR accounting for over 75 % of ADS hardware expense [16, 17].

In short, vision is sufficient for autonomy. Radar makes it resilient. LiDAR makes it expensive, but can still have its uses in training datasets as benchmarks. The proposed ideal multi-modal solution is outlined in Figure 7.

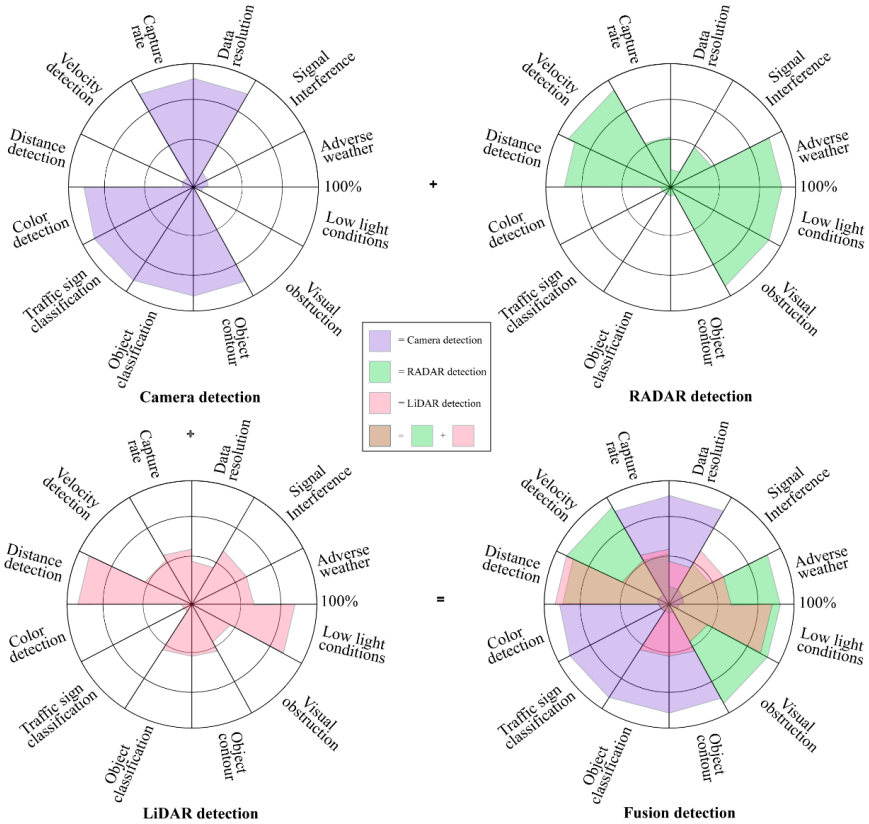

Fig. 7. A Multi-Modal Approach to ADS [13]

V. Conclusion

This study set out to evaluate whether camera-first autonomous driving systems can match or exceed the performance of LiDAR/radar fusion approaches. The answer is clear: for most conditions, they can—and they do so with greater efficiency, lower cost, and fewer integration challenges.

The evidence shows:

- Vision-primary pipelines achieve near-parity with fusion systems on 3D object detection and depth accuracy.
- Their energy and latency profiles outperform fusion systems by over 5x, enabling real-time inference on embedded hardware.
- Environmental failure modes, such as fog or night driving, can be mitigated with selective radar fallback, avoiding the need for costly LiDAR.
- End-to-end training on vision stacks reduces noise propagation and obviates sensor calibration.

These findings point to a clear path forward: deploy vision-first systems as the default, integrate radar for robustness where needed, and phase out LiDAR except in high-precision edge cases. With advances in simulation, world modeling, and vision transformers, the field now has the tools to deliver scalable autonomy without exotic sensors.

The next generation of autonomous vehicles will not be defined by how many sensors they carry, but by how intelligently they interpret the world with less. Vision, when empowered by the right models, is not just sufficient—it is optimal. [1]

[1] LLM was used strictly for formatting IAW IEEE Editorial Style Manual for Authors. All thoughts, prose, and conclusions are my own.

References

[1] SAE International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," Apr. 2021, Revised April 2021, Issued January 2014. DOI: 10.4271/J3016_202104. [Online]. Available: https://doi.org/10.4271/J3016_202104.

[2] S. Royo and M. Ballesta-Garcia, "An overview of lidar imaging systems for autonomous vehicles," *Applied Sciences*, vol. 9, no. 19, 2019, ISSN: 2076-3417. DOI: 10.3390/app9194093. [Online]. Available: https://www.mdpi.com/2076-3417/9/19/4093.

[3] Y. Zeng et al., "Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3434–3440, 2018. DOI: 10.1109/LRA.2018.2852843.

[4] Z. Han, X. Li, and Y. Zhou, "4d millimeter-wave radar in autonomous driving: A survey," *arXiv preprint arXiv:2306.04242*, 2023. [Online]. Available: https://arxiv.org/abs/2306.04242.

[5] Z. Li et al., "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," 2022. arXiv: 2203.17270 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2203.17270.

[6] K. Rana, L. Wang, and M. Gupta, "The perception systems used in fully automated vehicles: A comparative analysis," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 15 923–15 952, 2023. DOI: 10.1007/s11042-023-16000-4.

[7] Y. Chen, X. Huang, T. Ma, and et al., "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, Early access. DOI: 10.1109/TPAMI.2024.XXXXXXX.

[8] M. Goff et al., "Learning to drive from a world model," 2025. arXiv: 2504.19077 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2504.19077.

[9] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," 2021. arXiv: 2012.09841 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2012.09841.

[10] A. Vaswani et al., "Attention is all you need," 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[11] H. Haghighi, X. Wang, H. Jing, and M. Dianati, "Data-driven camera and lidar simulation models for autonomous driving: A review from generative models to volume renderers," *arXiv preprint arXiv:2402.10079*, 2024. [Online]. Available: https://arxiv.org/abs/2402.10079.

[12] D. Wu, F. Yang, B. Xu, P. Liao, and B. Liu, "A survey of deep learning based radar and vision fusion for 3d object detection in autonomous driving," 2024. arXiv: 2406.00714 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2406.00714.

[13] M. Hasanujjaman, M. Chowdhury, and Y. M. Jang, "Sensor fusion in autonomous vehicle with traffic surveillance camera system: Detection, localization, and ai networking," *Sensors*, vol. 23, p. 3335, Mar. 2023. DOI: 10.3390/s23063335.

[14] Q. Zhang, Y. Wu, and Z. Liu, "Multi-sensor fusion object detection in autonomous driving: A survey," *Sensors*, vol. 25, no. 9, p. 2794, 2023. DOI: 10.3390/s25092794.

[15] M. Dreissig, D. Scheuble, F. Piewak, and J. Boedecker, "Survey on lidar perception in adverse weather conditions," pp. 1–8, 2023. DOI: 10.1109/IV55152.2023.10186539.

[16] P. Wang, "Research on comparison of lidar and camera in autonomous driving," *Journal of Physics: Conference Series*, vol. 2093, p. 012 032, Nov. 2021. DOI: 10.1088/1742-6596/2093/1/012032.

[17] S. Sajjad, A. Hussain, and F. Alam, "A comparative analysis of camera, lidar and fusion-based deep neural networks for vehicle detection," *International Journal of Innovations in Science and Technology*, vol. 3, no. Special Issue, pp. 15–24, 2021. DOI: 10.33411/IJIST/2021030514. [Online]. Available: https://journal.50sea.com/index.php/IJIST/article/view/131.