Efficient Multimodal Geometry for Road-Scene Perception: Curb Detection and Zero-Shot Depth

Abstract

Road-Scene Geometry has become a test of how well researchers can connect performance with evidence quality, resource limits, and transfer across settings. This article critically maps combining cross-sensor attention and efficient state-space decoding for transferable geometric perception. The source set brings together 2 focal papers with 13 independently retrieved publications whose DOI or publisher records were checked before inclusion. The analysis is organized around sensor fusion, depth priors, attention efficiency, domain transfer, and road safety. To avoid reading outcomes from different settings as equivalent, the review compares research framing, method assumptions, and test envelope. Across the literature, the evidence indicates that advances in road-scene geometry become credible when representation, objective, and evaluation protocol are evaluated together and when uncertainty about distribution shift is reported explicitly. This organization relates method selection to decision risk and exposes recurring transfer threats, and proposes a research agenda centered on clear comparators, adversarial conditions, and inspectable evidence.

Keywords

Road-Scene Geometry; Sensor Fusion; Depth Priors; Attention Efficiency; Domain Transfer; Road Safety

1. Introduction

Work in road-scene geometry links scientific ambition and practical constraint. Progress requires not only stronger nominal performance but also explanations for when and why a method works. This difficulty is evident in combining cross-sensor attention and efficient state-space decoding for transferable geometric perception. An advantage observed in one composition, data source, cohort, location, or demand pattern can erode beyond the conditions that produced it. The comparison accordingly needs to preserve the unit of analysis and the decision context. A second task is to distinguish a mechanism from a correlation, a benchmark advantage from a deployment advantage, and an interpretable diagnostic from a causal explanation.

The evidence is examined through five lenses: sensor fusion, depth priors, attention efficiency, domain transfer, road safety. These dimensions matter because they jointly cover what changes, how change is observed, and which conditions must persist. The review is anchored in representation, objective, and evaluation protocol. Under that principle, innovation must be accompanied by validation; the comparison also requires aligned controls, explicit uncertainty, and credible resource or safety assumptions. The work reported here is a critical review and introduces no new experiment, cohort, measurement campaign, or benchmark score.

2. Methodological Foundations

A tractable account separates the input evidence, the transformation applied to it, the output used for evaluation, and the decision that follows. In road-scene geometry, the stages are frequently studied apart even though errors propagate across them. Bias in measurement can be carried forward and enlarged by the analytical method; an apparently accurate output can still be poorly calibrated for the final decision. This is why, ablation evidence should accompany aggregate performance. A useful evaluation makes explicit controlled assumptions alongside sources of variation, then selects metrics that correspond to the intended use rather than to convenience alone.

Boundary awareness provides the next principle. Sensor fusion defines the local design problem, while road safety determines whether the design can survive outside that boundary. Connecting the endpoints are depth priors and attention efficiency connect those ends by exposing trade-offs that a single score conceals. Here, adverse outcomes and perturbation analysis reveals the interval in which an explanation can still be defended. For applied work, documentation of deployment constraints is therefore part of the scientific result, not an administrative afterthought.

3. Comparative Evaluation

The target study X2-01, Falcon: Fused attention for lidar-camera curb detection [1], addresses a concrete part of road-scene geometry through attention-based fusion of LiDAR and camera cues for curb detection. It identifies a representation, mechanism, or evaluation boundary needed to compare claims. Against the focus on combining cross-sensor attention and efficient state-space decoding for transferable geometric perception, the paper is treated as a bounded design case: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis maintains the declared scope and triangulates the underlying assumptions.

The target study X2-02, WaveSamba: A Wavelet Transform SSM Zero-Shot Depth Estimation Decoder [2], addresses a concrete part of road-scene geometry through a wavelet-transform state-space decoder for zero-shot depth estimation. It identifies a representation, mechanism, or evaluation boundary needed to compare claims. Against the focus on combining cross-sensor attention and efficient state-space decoding for transferable geometric perception, the paper is treated as a bounded design case: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis maintains the declared scope and triangulates the underlying assumptions.

Across the focal studies, sensor fusion should be interpreted jointly with depth priors. A design can improve one but transfer cost into the coupled dimension. The design space can be represented as a matrix whose rows are operating conditions and whose columns are evidence quality, computation or process cost, robustness, and consequence of failure. The grid keeps an attractive mechanism from being detached from the circumstances that support it. The structure shows which ablations are informative: a component ablation should probe a declared mechanism instead of adding an isolated metric.

Attention efficiency links technical design with operational choice. The evaluation protocol needs to disaggregate routine and adverse conditions while making variability visible, and test plausible alternative explanations. Where distribution shift is central, calibration or sensitivity is as important as ranking. Where costs or hazards are asymmetric, utility-weighted assessment is more revealing than undifferentiated accuracy. These decisions need not homogenize studies; they make the reasons for their differences auditable.

4. Architecture and Learning Objectives

For triangulation, the literature includes Robust Curb Detection with Fusion of 3D-Lidar and Camera Data [3]; MF-BEVFusion: multiscale depth estimation and fully dynamic fusion for camera-LiDAR BEV 3D object detect... [4]; Dense Depth Map Estimation Based on Camera–LiDAR Sensor Fusion [5]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of road-scene geometry. Together they illuminate sensor fusion: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The design space is broadened by Zero-Shot Polarization-Intensity Physical Fusion Monocular Depth Estimation for High Dynamic Range Scenes [6]; DepthFusion: Depth-Aware Hybrid Feature Fusion for LiDAR-Camera 3D Object Detection [7]; Probabilistic multi-modal depth estimation based on camera–LiDAR sensor fusion [8]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of road-scene geometry. Together they illuminate depth priors: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

A first comparison set connects Obstacle detection for intelligent robots based on the fusion of 2D lidar and depth camera [9]; Real-Time Object Detection and Distance Measurement Enhanced with Semantic 3D Depth Sensing Using Camera... [10]; A LiDAR-depth camera information fusion method for human robot collaboration environment [11]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of road-scene geometry. Together they illuminate attention efficiency: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The neighboring evidence base brings together Obstacle detection for intelligent robots based on the fusion of 2D lidar and depth camera [12]; Adaptive Active Fusion of Camera and Single-Point LiDAR for Depth Estimation [13]; Sparse LiDAR and Stereo Fusion (SLS-Fusion) for Depth Estimation and 3D Object Detection [14]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of road-scene geometry. Together they illuminate domain transfer: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

A complementary strand is visible in Dense Depth-Map Estimation Based on Fusion of Event Camera and Sparse LiDAR [15]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of road-scene geometry. Together they illuminate road safety: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

5. Deployment and Governance

A practical workflow has five stages. First, define the decision and its failure cost. Second, specify the evidence and operating envelope. Third, choose a method whose inductive assumptions match that evidence. Fourth, test domain transfer and road safety under controlled perturbations. Finally, retain provenance so that an unexpected outcome can be traced to data, configuration, material batch, environmental condition, or user interaction. These stages keep combining cross-sensor attention and efficient state-space decoding for transferable geometric perception connected to observable requirements.

The cross-cutting implication is that responsible progress in road-scene geometry requires careful interfaces, not just strong components. Interfaces determine how uncertainty moves between measurement and inference, between algorithm and operator, or between microscopic design and system behavior. Teams spanning disciplines should predefine acceptance rules and triggers for revising conclusions. When those agreements are explicit, optimization can pursue efficiency without silently relaxing validity, safety, or interpretability.

6. Limitations and Future Directions

Several limitations qualify the synthesis, including divergent terms, methods, and documentation depth. A valid DOI record authenticates bibliographic facts without independently certifying empirical conclusions. Versioning is another source of uncertainty. Accordingly, focal Scholar strings remain intact and anomalous records receive separate comparison notes.

Subsequent studies need to preregister comparison protocols where feasible, make configurations computable and evaluate one meaningful change in context in deployment constraints. Research should also distinguish mechanistic failure classes rather than reporting totals only. For road-scene geometry, a valuable shared benchmark would combine ordinary performance, operational burden, uncertainty calibration, and resilience. Such a benchmark would reward designs that remain useful when evidence is incomplete or conditions change.

7. Conclusion

Scholarship in road-scene geometry forms an evidence network rather than a list of results rather than a collection of isolated scores. The focal studies show how combining cross-sensor attention and efficient state-space decoding for transferable geometric perception; the additional literature reveals the assumptions required to interpret those contributions. Across sensor fusion, depth priors, attention efficiency, domain transfer, and road safety, the recurring requirement is alignment: the representation or mechanism must match the problem, the evaluation must match the decision, and the deployment claim must match the tested boundary. Alignment makes transfer more cautious, replication more feasible, and failure more instructive.

References

1. Chen, Y., Long, Z., Wu, Y., & Chen, L. (2024). Falcon: Fused attention for lidar-camera curb detection.

2. Hou, B., & Long, Z. (2024). WaveSamba: A Wavelet Transform SSM Zero-Shot Depth Estimation Decoder. In 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (pp. 738-744). IEEE.

3. Tan, J., Li, J., An, X., & He, H. (2014). Robust Curb Detection with Fusion of 3D-Lidar and Camera Data. Sensors, 14(5), 9046-9073. https://doi.org/10.3390/s140509046
4. Liu, J., Yue, S., Hao, W., & Cai, Y. (2026). MF-BEVFusion: multiscale depth estimation and fully dynamic fusion for camera-LiDAR BEV 3D object detection. Journal of Electronic Imaging, 35(02). https://doi.org/10.1117/1.jei.35.2.023009
5. Bong, E. J., & Kee, S. C. (2026). Dense Depth Map Estimation Based on Camera–LiDAR Sensor Fusion. IEEE Sensors Journal, 26(4), 5891-5901. https://doi.org/10.1109/jsen.2025.3649237
6. Rao, R., Ouyang, Z., Chen, S., Chen, L., Huang, G., & Cui, C. (2026). Zero-Shot Polarization-Intensity Physical Fusion Monocular Depth Estimation for High Dynamic Range Scenes. Photonics, 13(3), 268. https://doi.org/10.3390/photonics13030268
7. Ji, M., Yang, J., & Zhang, S. (2026). DepthFusion: Depth-Aware Hybrid Feature Fusion for LiDAR-Camera 3D Object Detection. IEEE Transactions on Multimedia, 28, 7217-7227. https://doi.org/10.1109/tmm.2026.3668596
8. Obando-Ceron, J. S., Romero-Cano, V., & Monteiro, S. (2023). Probabilistic multi-modal depth estimation based on camera–LiDAR sensor fusion. Machine Vision and Applications, 34(5). https://doi.org/10.1007/s00138-023-01426-x
9. Meng, L., Zhao, H., & Fan, B. (2024). Obstacle detection for intelligent robots based on the fusion of 2D lidar and depth camera. International Journal of Hydromechatronics, 7(1). https://doi.org/10.1504/ijhm.2024.10060856
10. Yildiz, A. S., Meng, H., & Swash, M. R. (2025). Real-Time Object Detection and Distance Measurement Enhanced with Semantic 3D Depth Sensing Using Camera–LiDAR Fusion. Applied Sciences, 15(10), 5543. https://doi.org/10.3390/app15105543
11. Wang, Z., Li, P., Zhang, Q., Zhu, L., & Tian, W. (2025). A LiDAR-depth camera information fusion method for human robot collaboration environment. Information Fusion, 114, 102717. https://doi.org/10.1016/j.inffus.2024.102717
12. Fan, B., Zhao, H., & Meng, L. (2024). Obstacle detection for intelligent robots based on the fusion of 2D lidar and depth camera. International Journal of Hydromechatronics, 7(1), 67-88. https://doi.org/10.1504/ijhm.2024.135994
13. Tran, D. M., Ahlgren, N., Depcik, C., & He, H. (2023). Adaptive Active Fusion of Camera and Single-Point LiDAR for Depth Estimation. IEEE Transactions on Instrumentation and Measurement, 72, 1-9. https://doi.org/10.1109/tim.2023.3284129
14. Mai, N. A. M., Duthon, P., Khoudour, L., Crouzil, A., & Velastin, S. A. (2021). Sparse LiDAR and Stereo Fusion (SLS-Fusion) for Depth Estimation and 3D Object Detection. IET Conference Proceedings, 2021(1), 150-156. https://doi.org/10.1049/icp.2021.1442
15. Cui, M., Zhu, Y., Liu, Y., Liu, Y., Chen, G., & Huang, K. (2022). Dense Depth-Map Estimation Based on Fusion of Event Camera and Sparse LiDAR. IEEE Transactions on Instrumentation and Measurement, 71, 1-11. https://doi.org/10.1109/tim.2022.3144229