PEER-REVIEWED JOURNAL OF COMPUTER SCIENCE (PRJCS)

ISSN Online : Open Access Scholarly Journal

# Autonomous Vehicle Perception: Deep Learning Challenges and Solutions

Juby George

Assistant Professor, Department of Computer Applications, Marian College Kuttikkanam Autonomous, Kerala, India

**Article information**

Received: 5th March 2026
Received in revised form: 25th March 2026
Accepted: 28th March 2026
Available online: 9th April 2026

Volume: 1
Issue: 4
DOI: https://doi.org/10.5281/zenodo.19467460

**Abstract**

*Autonomous vehicles (AVs) depend on robust perception systems to interpret complex driving environments in real time. Deep learning has emerged as the dominant paradigm for AV perception, enabling breakthroughs in object detection, semantic segmentation, and sensor fusion. However, significant challenges persist, including adverse weather degradation, domain shift across geographic regions, computational constraints for real-time inference, and the long-tail distribution of rare but safety-critical scenarios. This article presents a comprehensive survey of deep learning approaches for autonomous vehicle perception, examining architectures for camera-based, LiDAR-based, and multi-modal fusion systems. We systematically analyze benchmark datasets including KITTI, nuScenes, and Waymo Open, evaluate state-of-the-art models across detection and segmentation tasks, and identify persistent gaps between research performance and deployment requirements. Our analysis reveals that while transformer-based architectures and self-supervised pre-training have substantially improved perception accuracy, challenges in real-time multi-sensor fusion, corner-case handling, and certification for safety-critical deployment remain open research problems requiring interdisciplinary solutions.*

**Keywords**: - autonomous vehicles, deep learning, object detection, LiDAR point cloud, sensor fusion, semantic segmentation, perception systems, KITTI benchmark

## I. INTRODUCTION

The pursuit of fully autonomous driving represents one of the most ambitious engineering challenges of the twenty-first century. At the core of every autonomous vehicle lies its perception system—the computational framework responsible for transforming raw sensor data into a coherent understanding of the surrounding environment [1]. This perception capability must reliably detect and classify other vehicles, pedestrians, cyclists, traffic signs, lane markings, and myriad other objects under diverse conditions, all within stringent latency constraints measured in milliseconds [2].

Deep learning has fundamentally transformed AV perception over the past decade. Convolutional neural networks (CNNs) first demonstrated superhuman performance on image classification tasks with the introduction of ResNet [3], and this success rapidly propagated to object detection through architectures such as YOLO [4] and Faster R-CNN [5]. More recently, transformer-based models including DETR [6] and BEVFormer [7] have challenged the CNN paradigm by leveraging attention mechanisms for global context modeling. Simultaneously, point cloud processing networks such as PointNet [8] and PointPillars [9] have enabled direct 3D perception from LiDAR data, while multi-modal fusion approaches attempt to combine the complementary strengths of cameras, LiDAR, and radar [10].

Despite these advances, a significant gap persists between benchmark performance and real-world deployment requirements. Models that achieve state-of-the-art results on curated datasets frequently degrade under adverse weather, unusual lighting, or geographic domain shift [11]. The long-tail distribution of driving scenarios means that rare but

safety-critical events—such as a child running onto the road or an overturned vehicle—are underrepresented in training data [12]. Furthermore, the computational demands of high-accuracy models often conflict with the real-time processing requirements of safety-critical systems [13].

This article provides a comprehensive examination of deep learning for autonomous vehicle perception. We survey the sensor modalities employed in modern AV systems, analyze benchmark datasets and evaluation protocols, review state-of-the-art architectures for detection and segmentation, and critically assess the open challenges that must be addressed before widespread deployment becomes feasible. Our analysis spans both camera-based and LiDAR-based approaches, as well as multi-modal fusion strategies, providing a unified perspective on the current state and future directions of AV perception research.

## II. SENSOR MODALITIES FOR AUTONOMOUS VEHICLE PERCEPTION

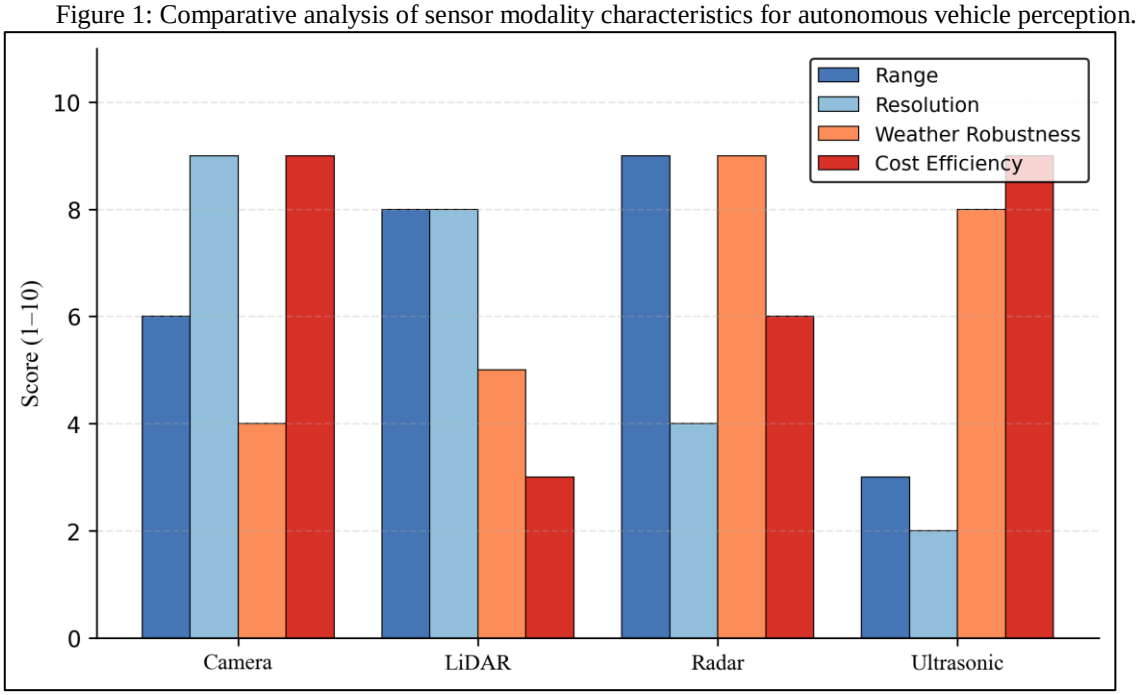
Autonomous vehicles employ a heterogeneous suite of sensors, each offering distinct advantages and limitations. Understanding these trade-offs is essential for designing effective perception systems. The four primary sensor modalities—cameras, LiDAR, radar, and ultrasonic sensors—differ substantially in range, resolution, weather robustness, and cost, as illustrated in Figure 1.

Cameras provide dense, high-resolution color imagery at low cost, making them indispensable for tasks requiring texture and color information such as traffic sign recognition and lane detection [14]. However, cameras are passive sensors susceptible to illumination variations, glare, and degradation in rain or fog. Monocular depth estimation remains fundamentally ill-posed, though stereo configurations and learned depth prediction have partially addressed this limitation [15].

LiDAR (Light Detection and Ranging) sensors emit laser pulses to generate precise 3D point clouds of the environment, providing accurate depth information independent of ambient lighting [2]. Modern spinning LiDAR units such as the Velodyne VLP-128 produce over 4 million points per second with centimeter-level accuracy at ranges exceeding 200 meters. However, LiDAR performance degrades in heavy rain, snow, and fog due to laser pulse scattering, and the sensors remain significantly more expensive than cameras [16].

Radar sensors offer superior weather robustness and long-range detection capabilities, functioning reliably in rain, fog, and dust conditions that degrade both cameras and LiDAR [17]. Automotive radar operates primarily at 77 GHz and provides velocity measurements via the Doppler effect. However, radar angular resolution is substantially lower than either cameras or LiDAR, limiting fine-grained object classification. Ultrasonic sensors complement the suite for short-range applications such as parking assistance, operating at ranges below 5 meters with high reliability but limited applicability for highway driving scenarios [18].

Figure 1: Comparative analysis of sensor modality characteristics for autonomous vehicle perception.




Scores are normalized on a 1–10 scale where higher values indicate superior performance. Cost efficiency reflects affordability (higher = more affordable).

## III. BENCHMARK DATASETS AND EVALUATION PROTOCOLS

The development of large-scale annotated datasets has been instrumental in advancing AV perception research. These benchmarks provide standardized evaluation protocols that enable fair comparison across methods and drive systematic progress. Table 1 summarizes the major autonomous driving datasets that have shaped the field.

The KITTI benchmark [2], introduced in 2012, was the first large-scale dataset to provide synchronized camera and LiDAR data with 3D bounding box annotations for autonomous driving research. Despite its relatively modest scale by contemporary standards, KITTI remains the most widely cited AV perception benchmark and established evaluation conventions—particularly the use of mean Average Precision (mAP) at IoU thresholds of 0.5 and 0.7—that persist across newer datasets. However, KITTI's geographic limitation to Karlsruhe, Germany, and its predominantly fair-weather conditions limit its representativeness [2].

The nuScenes dataset [19] substantially expanded the scope of AV benchmarking by providing 1,000 driving scenes captured in Boston and Singapore, annotated with 1.4 million 3D bounding boxes across 23 object classes. Its inclusion of tropical weather conditions and dense urban scenarios from two continents improved geographic diversity. The Waymo Open Dataset [20] further raised the bar with 1,150 scenes containing over 12 million 3D bounding box annotations, captured across multiple U.S. cities with diverse weather and lighting conditions. More recently, Argoverse 2 [21] and ONCE [22] have contributed additional geographic coverage and novel annotation types including motion forecasting labels and semi-supervised learning scenarios.

Table 1. Comparison of Major Autonomous Driving Datasets

| Dataset | Year | Sensors | Scenes | 3D Boxes | Country |
|---|---|---|---|---|---|
| KITTI [2] | 2012 | Camera, LiDAR, GPS | 22 seq | 80K | Germany |
| nuScenes [19] | 2020 | Camera, LiDAR, Radar | 1,000 | 1.4M | USA / Singapore |
| Waymo Open [20] | 2020 | Camera, LiDAR | 1,150 | 12M | USA |
| Argoverse 2 [21] | 2021 | Camera, LiDAR | 1,000 | 5M | USA |
| ONCE [22] | 2021 | Camera, LiDAR | 581 | 417K | China |

## IV. DEEP LEARNING ARCHITECTURES FOR AV PERCEPTION

### A. Camera-Based Object Detection

Camera-based object detection has evolved through two principal paradigms: single-stage and two-stage detectors. Two-stage approaches, exemplified by Faster R-CNN [5], first generate region proposals and then classify and refine each proposal independently. This decomposition typically yields higher accuracy but at increased computational cost. The Region Proposal Network (RPN) introduced in Faster R-CNN remains foundational, generating anchor-based proposals that are subsequently processed by a classification and regression head [5].

Single-stage detectors eliminate the proposal generation step, directly predicting bounding boxes and class probabilities from dense feature maps. YOLO (You Only Look Once) [4] pioneered this approach by formulating detection as a single regression problem, achieving real-time performance at the cost of reduced accuracy on small objects. Subsequent YOLO iterations progressively addressed these limitations: YOLOv4 [23] incorporated Cross-Stage Partial connections and Mish activation, while YOLOv7 [24] introduced Extended Efficient Layer Aggregation Networks (E-ELAN) achieving state-of-the-art speed-accuracy trade-offs. The SSD (Single Shot MultiBox Detector) [25] architecture contributed the multi-scale feature map strategy, detecting objects at different resolutions to improve small-object performance.

Transformer-based detectors represent the latest architectural evolution. DETR (DEtection TRansformer) [6] reformulated object detection as a direct set prediction problem, eliminating hand-crafted components such as non-maximum suppression and anchor generation. While the original DETR suffered from slow convergence and poor small-object detection, subsequent variants including Deformable DETR [26] and DINO [27] have largely resolved these issues. Figure 2 presents a comparative analysis of these detection architectures on the KITTI benchmark, demonstrating the performance evolution across model generations.

Figure 2: Object detection model performance on the KITTI benchmark

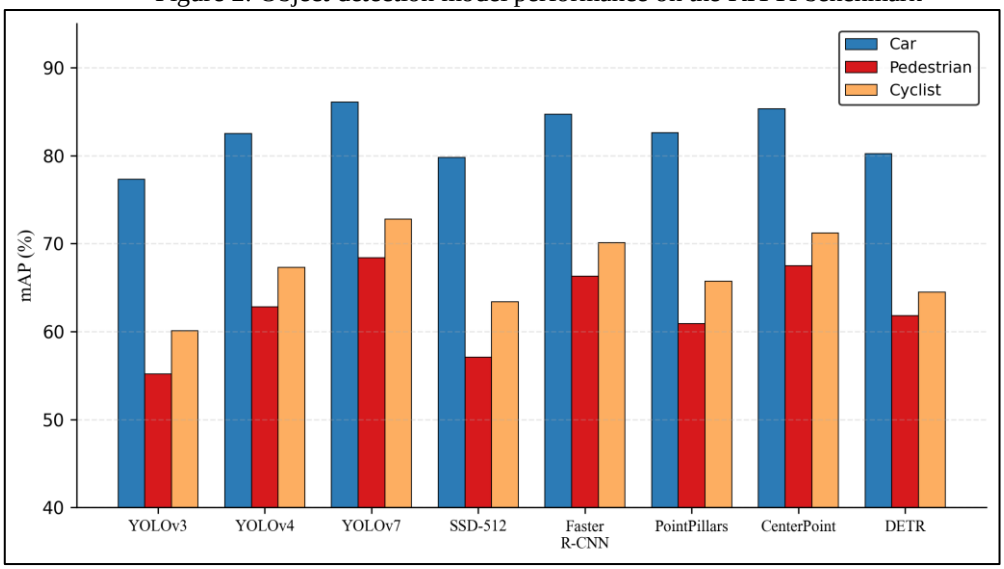



Figure 2. Object detection model performance on the KITTI benchmark across three categories (Car, Pedestrian, Cyclist) measured by mean Average Precision (mAP) at IoU = 0.7. YOLOv7 and CenterPoint achieve the highest scores across categories.

**B. LiDAR-Based 3D Object Detection**

Processing LiDAR point clouds for 3D object detection presents unique challenges due to the unstructured, sparse, and unordered nature of point cloud data. PointNet [8] addressed the permutation invariance challenge by applying symmetric functions (max pooling) to learn point-wise features, establishing the foundation for direct point cloud processing. PointNet++ [28] extended this framework with hierarchical feature learning using set abstraction layers to capture local geometric structures at multiple scales.

Voxel-based methods discretize the continuous 3D space into a regular grid, enabling the application of 3D convolutions. VoxelNet [29] introduced end-to-end voxel feature learning, while SECOND [30] dramatically improved efficiency through sparse convolutions that skip empty voxels. PointPillars [9] simplified the voxelization to vertical columns (pillars), enabling 2D convolution processing and achieving real-time performance exceeding 60 FPS on the KITTI benchmark. CenterPoint [31] adopted a center-based detection paradigm, representing objects by their center points rather than bounding box corners, achieving superior performance on the nuScenes and Waymo Open benchmarks.

**C. Multi-Modal Sensor Fusion**

Multi-modal fusion leverages the complementary strengths of different sensors to achieve more robust and accurate perception than any single modality alone. Fusion strategies are categorized by the processing stage at which modalities are combined: early fusion concatenates raw sensor data, mid-level fusion combines intermediate feature representations, and late fusion merges independent detection outputs [10]. Mid-level fusion has emerged as the predominant approach, exemplified by architectures such as MVXNet [32] and TransFusion [33], which align camera features with LiDAR point clouds in a shared representation space.

Bird's-eye view (BEV) representations have gained prominence as a unifying framework for multi-sensor fusion. BEVFusion [34] projects both camera and LiDAR features into a common BEV space, enabling efficient feature-level fusion. BEVFormer [7] leverages spatiotemporal transformers to aggregate multi-camera features into BEV representations without LiDAR, demonstrating that camera-only systems can approach LiDAR-based accuracy when provided with sufficient training data and architectural innovations.

Table 2. Deep Learning Architectures for Autonomous Vehicle Perception

| Model | Task | Input | mAP / mIoU | FPS | Year |
|---|---|---|---|---|---|
| Faster R-CNN [5] | 2D Detection | Camera | 84.7 mAP | 17 | 2015 |
| YOLOv7 [24] | 2D Detection | Camera | 86.1 mAP | 55 | 2022 |
| DETR [6] | 2D Detection | Camera | 80.2 mAP | 28 | 2020 |
| PointPillars [9] | 3D Detection | LiDAR | 82.6 mAP | 62 | 2019 |
| CenterPoint [31] | 3D Detection | LiDAR | 85.3 mAP | 30 | 2021 |
| PointNet++ [28] | Segmentation | LiDAR | 53.4 mIoU | 10 | 2017 |
| BEVFusion [34] | 3D Detection | Camera+LiDAR | 71.4 NDS | 15 | 2022 |
| BEVFormer [7] | 3D Detection | Camera | 56.9 NDS | 4 | 2022 |

Table 2 summarizes the principal deep learning architectures employed for AV perception, highlighting the trade-offs between accuracy (mAP or mIoU), inference speed (FPS), and input modality. Camera-based detectors such as YOLOv7 offer the highest frame rates, while LiDAR-based methods like CenterPoint achieve superior 3D detection accuracy. Multi-modal fusion approaches such as BEVFusion demonstrate that combining modalities yields the most robust performance, albeit at increased computational cost.

## V. PERSISTENT CHALLENGES IN AV PERCEPTION

### A. Adverse Weather and Lighting Conditions

Perception system performance degrades substantially under adverse environmental conditions. Rain, fog, and snow scatter LiDAR pulses and introduce noise in camera images, while direct sunlight and nighttime darkness challenge camera-based detection [11]. Empirical studies demonstrate that state-of-the-art detectors can lose 20–40% of their accuracy in heavy rain compared to clear conditions [35]. Domain adaptation and adverse-weather data augmentation techniques, including physics-based rain and fog simulation [36], represent active areas of research but have not yet closed this performance gap.

### B. Domain Shift and Geographic Generalization

Models trained on datasets from specific geographic regions frequently fail to generalize to new environments with different road layouts, traffic patterns, vehicle types, and driving behaviors [37]. This domain gap is particularly acute between Western and Asian driving environments, where traffic density, infrastructure design, and road user composition differ dramatically. Unsupervised domain adaptation techniques, including self-training and adversarial alignment, have shown promise but remain insufficient for safety-critical deployment [38].

### C. Long-Tail Distribution and Corner Cases

Autonomous driving scenarios follow a long-tail distribution where common situations (highway driving, regular intersections) dominate the training data while rare but critical events (construction zones, emergency vehicles, unusual obstacles) are severely underrepresented [12]. Standard training procedures bias models toward frequent patterns, resulting in poor performance on the rare events that matter most for safety. Strategies addressing this challenge include synthetic data generation through simulation [39], class-balanced sampling, and few-shot learning approaches for rare object detection [40].

### D. Real-Time Computation and Edge Deployment

Safety-critical perception requires inference latencies below 100 milliseconds, corresponding to frame rates of at least 10 FPS for the complete perception pipeline [13]. However, the most accurate models—particularly transformer-based architectures and multi-modal fusion systems—often exceed this latency budget on automotive-grade hardware. Model compression techniques including knowledge distillation [41], quantization, and neural architecture search offer pathways toward deployment-feasible efficiency, but accuracy-latency trade-offs remain a fundamental tension in AV perception system design.

### VI. Emerging Solutions and Future Directions

Several emerging research directions offer promising pathways toward addressing the persistent challenges in AV perception. Self-supervised and semi-supervised learning methods reduce dependence on expensive human annotations by leveraging the vast quantities of unlabeled driving data available from fleet operations [42]. Contrastive learning frameworks such as SimCLR and BYOL have demonstrated effective pre-training for visual representations that transfer well to downstream perception tasks [43].

Neural radiance fields (NeRF) and related 3D scene reconstruction techniques enable photorealistic novel-view synthesis from driving data, providing a powerful augmentation mechanism for generating diverse training scenarios from limited real-world captures [44]. World models that learn predictive representations of driving environments are being explored for simulation-based training and as components of end-to-end autonomous driving systems [45].

Occupancy networks represent an emerging alternative to traditional bounding-box-based perception, predicting dense 3D occupancy grids that capture the geometry of arbitrary-shaped objects and free space simultaneously [46]. This representation is particularly advantageous for detecting unusual objects that fall outside predefined category taxonomies. Additionally, vehicle-to-everything (V2X) communication enables cooperative perception by sharing sensor data across connected vehicles and infrastructure, extending the effective perception range beyond line-of-sight limitations [47].

Foundation models pre-trained on internet-scale data, including vision-language models such as CLIP [48], are beginning to influence AV perception through open-vocabulary detection and zero-shot transfer capabilities. These models offer the potential to recognize novel objects without explicit training examples, addressing the long-tail challenge. However, their computational requirements currently preclude real-time deployment, motivating research into efficient adaptation and distillation strategies.

## VII. CONCLUSION

Deep learning has driven transformative progress in autonomous vehicle perception, enabling detection and segmentation performance that was unimaginable a decade ago. As demonstrated in our analysis (Figure 1, Figure 2, Tables 1–2), modern architectures achieve impressive benchmark results across camera-based, LiDAR-based, and multi-modal fusion approaches. YOLOv7 and CenterPoint represent the current performance frontier for 2D and 3D detection respectively, while BEV-based fusion methods offer the most robust multi-sensor perception.

Nevertheless, substantial gaps remain between benchmark performance and the requirements of safe, reliable deployment. Adverse weather degradation, geographic domain shift, long-tail event coverage, and real-time computational constraints constitute the primary barriers to widespread autonomous vehicle deployment. Addressing these challenges will require advances not only in model architectures but also in training paradigms (self-supervised learning, simulation-based training), evaluation protocols (safety-focused benchmarks, corner-case testing), and system-level design (sensor redundancy, graceful degradation, V2X cooperation).

Future research should prioritize closing the sim-to-real gap, developing robust perception under degraded sensor conditions, and establishing rigorous certification frameworks that can provide statistical guarantees on perception system reliability. The convergence of foundation models, occupancy-based representations, and cooperative perception through V2X communication promises to define the next generation of autonomous vehicle perception systems, ultimately enabling safe and equitable autonomous mobility.

## REFERENCES


[1] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, 2020.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3354–3361.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[7] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–18.

[8] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 652–660.

[9] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 12697–12705.

[10] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1907–1915.

[11] M. Bijelic, T. Gruber, F. Mannan, D. Kraus, W. Ritter, K. Dietmayer, and H. P. A. Lensch, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11682–11692.

[12] K. Wong, S. Wang, M. Ren, M. Liang, and R. Urtasun, "Identifying unknown instances for autonomous driving," in *Proc. Conf. Robot Learning (CoRL)*, 2020, pp. 384–393.

[13] S. Liu, L. Li, J. Tang, S. Wu, and J. L. Gaudiot, "Creating autonomous vehicle systems," *Synth. Lect. Comput. Sci.*, vol. 8, no. 2, pp. 1–186, 2020.

[14] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1484–1497, 2012.

[15] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3828–3838.

[16] R. Roriz, J. Cabral, and T. Gomes, "Automotive LiDAR technology: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6282–6297, 2022.

[17] F. Engels, P. Heidenreich, M. Wintermantel, L. Stacker, M. Al Kadi, and A. M. Zoubir, "Automotive radar signal processing: Research directions and practical challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 865–878, 2021.

[18] J. Borenstein and Y. Koren, "Real-time obstacle avoidance for fast mobile robots in cluttered environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 1990, pp. 572–577.

[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11621–11631.

[20] P. Sun *et al*., "Scalability in perception for autonomous driving: Waymo Open Dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2446–2454.

[21] B. Wilson, W. Qi, T. Aber, J. Robinson, S. Meng, B. Fox, S. Rami, and J. Hays, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Proc. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks*, 2021, pp. 1–14.

[22] J. Mao *et al*., "One million scenes for autonomous driving: ONCE dataset," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–14.

[23] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[24] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7464–7475.
[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.
[26] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021, pp. 1–16.
[27] H. Zhang *et al*., "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023, pp. 1–18.
[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5099–5108.
[29] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4490–4499.
[30] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.
[31] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11784–11793.
[32] V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-Net: Multimodal VoxelNet for 3D object detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 7276–7282.
[33] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C. L. Tai, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1090–1099.
[34] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 2774–2781.
[35] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *Int. J. Comput. Vis.*, vol. 126, no. 9, pp. 973–992, 2018.
[36] M. Tremblay, S. S. Halder, R. de Charette, and J. P. Lalonde, "Rain rendering for evaluating and improving robustness to bad weather," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 341–360, 2021.
[37] X. Wang, M. Cai, F. Sohel, N. Anwar, and R. Dobson, "Adversarial generation of training examples: Applications to moving vehicle license plate recognition," *arXiv preprint arXiv:1707.03124*, 2017.
[38] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive Faster R-CNN for object detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3339–3348.
[39] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Conf. Robot Learning (CoRL)*, 2017, pp. 1–16.
[40] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, 2019.
[41] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[42] S. Chen, B. Zheng, S. Hilliges, and L. Van Gool, "Efficient and robust 2D-to-BEV representation learning via geometry-guided kernel transformer," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023, pp. 1–15.
[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2020, pp. 1597–1607.
[44] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2022.
[45] D. Hu, Z. Chen, and D. Zhao, "World models for autonomous driving: An initial survey," *IEEE Trans. Intell. Veh.*, vol. 8, no. 10, pp. 4288–4295, 2023.
[46] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 21729–21740.
[47] R. Xu, H. Xiang, Z. Tu, X. Xia, M. H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 107–124.
[48] A. Radford *et al*., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2021, pp. 8748–8763.