Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD37798

Camera–LiDAR–Radar Multi-sensor Fusion for Autonomous Driving Perception: A Survey

Yanze Li

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
hctvqvq87@outlook.com

Abstract. Camera, LiDAR, and radar are core sensing modalities of autonomous driving perception. Cameras provide dense appearance but inferred depth, LiDAR provides accurate 3D geometry but does not directly measure velocity and degrades in rain, fog, or dust, radar is relatively robust to adverse weather and measures Doppler velocity, yet returns sparse, noisy points. Fusion can therefore raise the accuracy and reliability of 3D detection, tracking, segmentation, and localization. A structured review of the 2019–2025 literature yields four findings. First, bird's-eye-view representation has emerged as a dominant fusion backbone because it removes the camera–LiDAR viewport mismatch that limited earlier proposal-level and ROI-pooling designs. Second, radar remains underused: the bottleneck often lies in its sparse and noisy point cloud rather than in the fusion operator itself. Third, spatial–temporal calibration and adverse-weather degradation set the practical performance ceiling more often than the fusion architecture itself. Fourth, the field is moving toward end-to-end and foundation-model paradigms. The survey organizes methods with a three-view taxonomy—processing stage, representation, and sensor pair—and maps these findings onto open problems and near-term research directions.

Keywords: Multi-sensor fusion, autonomous driving, camera–LiDAR–radar, bird's-eye view, 3D object detection

1. Introduction

Autonomous driving perception has evolved from single-sensor pipelines to tightly coupled multi-sensor fusion [1]. Camera-only systems offer dense semantics at low cost but fail to recover metric depth reliably in weak-texture, low-light, or glare conditions. LiDAR-only systems provide accurate 3D geometry and remain a core backbone for 3D detectors, yet performance degrades in rain, fog, and dust, and LiDAR cannot directly measure radial velocity [2, 3]. Radar is robust in adverse weather and directly measures Doppler velocity, but its returns are sparse and difficult to align with object boundaries [4]. No single modality covers the full perception envelope, motivating fusion (Figure 1). This complementarity reveals three persistent gaps. Methodologically, fusion designs are fragmented across processing stage (early/feature/late), representation (BEV vs. perspective), and sensor pair, with few taxonomies treating these dimensions as orthogonal. Second, system-level: calibration and synchronization impose a silent ceiling, even minor extrinsic errors or temporal skews severely degrade otherwise strong architectures. Third, practically, it remains unclear which

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD37798

modality combinations justify their calibration and association overhead and how fusion should respond when one sensor fails or degrades.

This survey addresses those gaps with a structured review of perception-level fusion (2019-2025), together with the earlier works that define each sub-area. Methods are organized by a three-view taxonomy—processing stage, representation, and sensor pair—and compared using results reported on public benchmarks [5, 6]. The aim is a decision guide for practitioners, not a new algorithm. Three questions structure the paper: (1) which fusion design best matches each perception task, (2) which modality combinations justify their calibration cost, and (3) how fusion should adapt when one sensor degrades. Multi-sensor fusion is essential for autonomous driving because no single modality delivers reliable perception across all driving conditions. By combining the complementary strengths of camera, LiDAR, and radar, fusion-based systems achieve robust 3D detection, tracking, and scene understanding—capabilities critical to vehicle safety and scalability.

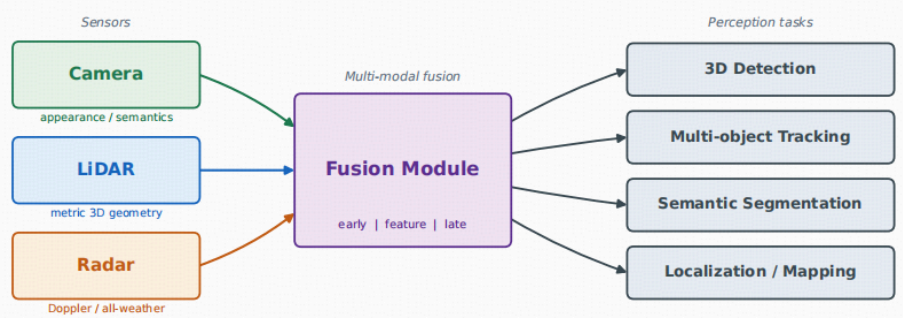


Figure 1. Camera, LiDAR, and radar feed a fusion module for autonomous-driving perception

2. Sensor characteristics and complementarity

Cameras are passive imagers that deliver dense color, texture, and appearance at low hardware cost, but depth is inferred rather than measured and fails under weak texture, low illumination, or strong glare [1]. LiDAR is an active time-of-flight ranger that yields accurate 3D geometry, its failure modes are weather-dependent attenuation and the absence of native velocity [2, 3]. Radar uses millimeter-wave Doppler to measure radial velocity and remains robust under rain, fog, and dust, but angular resolution is coarse and returns are sparse, so semantic detail is weak [4].

Table 1. Qualitative comparison of the three perception modalities

Property	Camera	LiDAR	Radar
Depth	inferred, unreliable	accurate	measured, coarse
Semantics	dense	limited	very limited
Velocity	indirect (temporal)	none	direct (Doppler)
Weather	glare-sensitive	rain/fog-sensitive	robust
Resolution	high (2D)	medium (3D)	low (sparse)
Cost	low	high	medium

The complementarity is principled. Camera supplies dense semantic information, LiDAR supplies accurate metric geometry, and radar supplies direct Doppler velocity and greater robustness under adverse weather. Fusion is most useful when failure modes are sensor-specific rather than scene-specific—the usual regime in autonomous driving (See Table 1). Adverse weather makes the argument concrete. Bijelic et al. show that deep fusion of camera, LiDAR, and radar substantially

DOI: 10.54254/2755-2721/2026.LD37798

outperforms a LiDAR-only baseline in unseen fog and rain, because radar and camera remain informative when LiDAR is attenuated. Fusion is most valuable exactly where a single modality degrades [7].

3. Spatial–temporal calibration and synchronization

Fusion is only as good as the geometric and temporal alignment among sensors. We separate three concerns and, more importantly, how calibration sensitivity varies by fusion stage. Spatial calibration estimates the intrinsic and extrinsic transforms [8]. Target-based methods use calibration boards or reflective spheres, target-less methods exploit natural edges or motion. The image–point-cloud projection that underpins camera–LiDAR fusion is especially brittle: even a small rotational error can produce substantial projection shifts at long range (Figure 2).

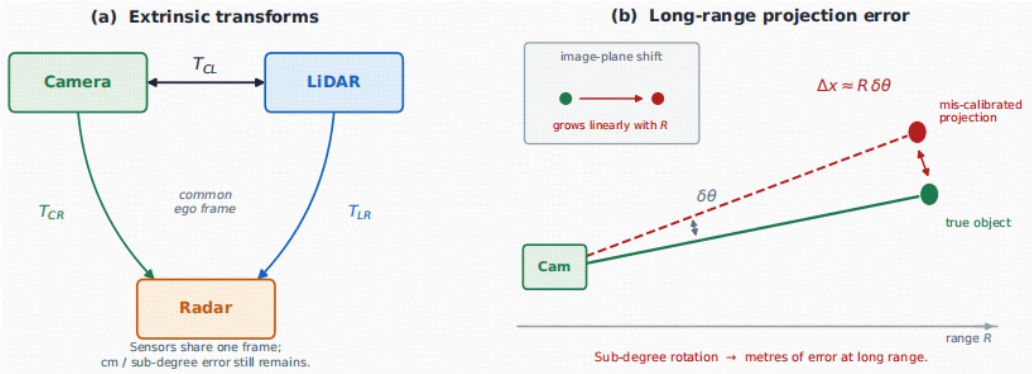


Figure 2. Extrinsic transforms among camera, LiDAR, and radar, and the growth of projection error with range

Temporal synchronization aligns timestamps by hardware triggering or software estimation and must compensate for rolling shutter and ego-motion so that measurements refer to a common instant [8]. Unsynchronized streams force interpolation or motion compensation, both of which inject error. The impact on fusion is stage-dependent. Data-level (early) fusion focuses raw or projected measurements and is the most sensitive: a mis-calibrated projection scatters valid geometry onto the wrong pixels. Feature-level fusion is moderately sensitive, as learned representations may tolerate small misalignments. Decision-level (late) fusion associates independent per-sensor outputs and is more robust to cross-sensor misalignment, at the cost of discarding fine-grained complementary signal. The stage extracting the most complementarity is therefore also the most vulnerable to calibration error. The focus here is this dependence, not a tutorial on calibration algorithms.

4. Taxonomy of fusion methods

Fusion methods are organized along three orthogonal views—processing stage, representation and fusion architecture, and sensor pair. The views are not hierarchical: a BEV transformer may operate at the feature level between a specific sensor pair.

4.1. Fusion by processing stage

Data-level (early) fusion combines raw or projected measurements before detection. PointPainting projects LiDAR points into image segmentation masks and appends the semantic features to each

DOI: 10.54254/2755-2721/2026.LD37798

point, an input-level concatenation that keeps the raw point cloud intact [9]. This stage preserves the richest signal but demands accurate calibration and high bandwidth (Figure 3).

Feature-level fusion interacts at intermediate representations: image-to-point pooling, cross-attention between camera and LiDAR features. Multi-task architectures may further share a fused backbone across perception heads. MV3D is an earlier region/proposal-based fusion approach that fuses multi-view regions of interest [10]. Liang et al. use depth completion as an auxiliary feature task in multi-task fusion [11]. DeepFusion applies gated fusion between camera and LiDAR features to exploit cross-modal cues while remaining robust to modest mis-calibration [12]. This approach has become widely used because it balances complementarity information against calibration sensitivity. Decision-level (late) fusion integrates per-sensor outputs through track association or track-level fusion [1]. A representative detection-level instance is CLOCs, which fuses camera and LiDAR detection candidates before non-maximum suppression on KITTI [13]. It is robust and modular but loses fine-grained cues and can suffer from conflicting detections.

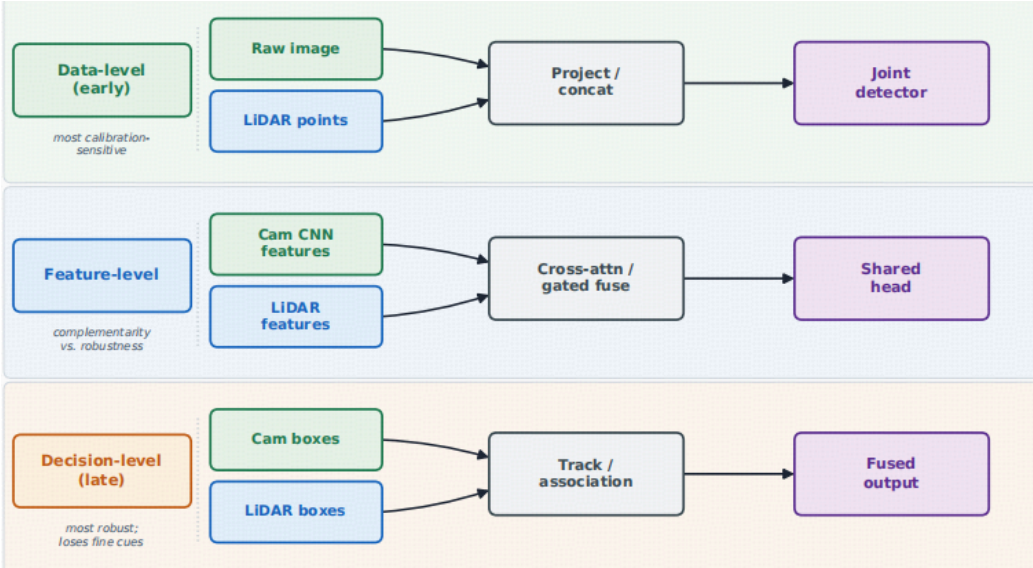


Figure 3. Three processing-stage fusion paradigms. The rows represent alternative fusion strategies rather than a sequential pipeline

4.2. Representation and fusion architecture

Representation and fusion architecture are orthogonal to the processing stage: a BEV or transformer-based design may be integrated at different stages of the fusion pipeline. This dimension is treated separately because it has become a dominant design paradigm in recent fusion research. BEV-based fusion uses a shared BEV grid to represent multi-sensor information. It reduces the camera–LiDAR viewport mismatch that limited earlier perspective-view fusion designs. Lift-Splat-Shoot (LSS) established explicit lift-to-BEV for camera-based perception, BEVFormer introduced spatiotemporal BEV queries from multi-camera images, and BEVFusion extended BEV representation to multi-task, multi-sensor camera-LiDAR perception [14-16] (Figure 4). BEV is now the de facto backbone of detection and segmentation stacks.

DOI: 10.54254/2755-2721/2026.LD37798

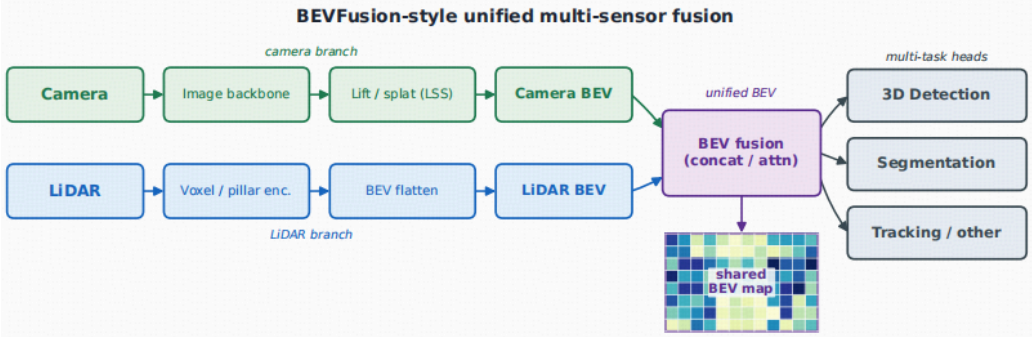


Figure 4. BEVFusion-style architecture: camera and LiDAR branches meet in a shared BEV map that feeds multi-task heads

Transformer-based fusion uses attention to align and integrate modalities without relying entirely on hand-crafted projection [17]. TransFusion applies camera-to-LiDAR cross-attention for 3D proposals, while FUTR3D uses a unified query-based detector across camera, LiDAR, and radar [18]. These two paradigms currently define the state of the art.

4.3. Sensor-pair and three-modality fusion

Camera–LiDAR fusion has the densest literature because it combines image-based appearance and semantic information with accurate 3D geometry. PointPainting and BEVFusion are representative approaches [9, 16]. Camera–radar fusion adds velocity at low cost, CenterFusion injects radar detections into a camera-based 3D detector [19]. LiDAR–radar fusion complements range with velocity under weather, dedicated detectors such as RadarNet and LiRaFusion begin to close the gap left by camera–LiDAR-centric work [2, 4, 20, 21]. Camera–LiDAR–radar fusion is the full three-modality setting, FUTR3D is a representative query-based approach [18]. Published three-modality work remains comparatively scarce, with the additional calibration and association cost representing an important barrier to adoption. Table 2 summarizes the representative methods discussed above along the three taxonomy views.

Table 2. Summary of major fusion methods and architectures across the three taxonomy views

Method	Stage	Representation	Sensors	Dataset	Note
PointPainting	data-level	Point features	Camera+LiDAR	nuScenes	input-level
MV3D	feature-level	BEV/FV/image ROI	Camera+LiDAR	KITTI	proposal-level
MMF	feature-level	BEV (depth-aux)	Camera+LiDAR	KITTI	multi-task
DeepFusion	feature-level	Gated features	Camera+LiDAR	Waymo	gated fusion
BEVFusion	feature-level	BEV	Camera+LiDAR	nuScenes	unified BEV
TransFusion	feature-level	BEV (cross-attn)	Camera+LiDAR	nuScenes	cross-attention
FUTR3D	feature-level	Query-based	Camera+LiDAR+Radar	nuScenes	unified query
CenterFusion	feature-level	BEV (center)	Camera+Radar	nuScenes	radar-injected
UniAD	—	BEV	Camera-only	nuScenes	end-to-end (ref.)
RadarNet	data-level	Voxel/BEV	LiDAR+Radar	nuScenes	early+late hybrid
LiRaFusion	feature-level	Voxel/BEV	LiDAR+Radar	nuScenes	early+middle hybrid
CLOCs	decision-level	Detection candidates	Camera+LiDAR	KITTI	detection-level (late)

DOI: 10.54254/2755-2721/2026.LD37798

Cross-paper detection scores (e.g., mAP/NDS) are not directly comparable because datasets, splits, backbones, and test settings differ, Table 2 is therefore intended as a qualitative comparison. All three processing stages are represented, while CLOCs illustrates decision-level fusion. UniAD is camera-only and is included as a reference architecture rather than as a fusion method.

5. Applications

Perception tasks impose different demands on sensors and fusion. 3D object detection. BEV feature fusion dominates as it removes view mismatch [16]. LiDAR-only baselines remain competitive at short range with clean geometry, but fusion becomes particularly valuable at long range and under occlusion: radar velocity helps resolve motion ambiguity, and camera appearance helps recover truncated shapes that LiDAR sparsity misses. TransFusion and FUTR3D show consistent gains from the complementary modality on nuScenes [17, 18]. Multi-object tracking. Radar Doppler is particularly useful for association under occlusion and for initializing tracks before appearance matures [1, 4]. Late-fusion trackers exploit this by associating radar tracks with camera or LiDAR tracks. Semantic segmentation and drivable-area estimation. Camera semantics complement LiDAR geometry, BEV fusion allows these complementary cues to contribute to a common spatial representation [16]. Localization and mapping. Camera–LiDAR fusion can improve localization robustness and reduce drift, while radar is still rarely the primary localization cue [5, 8].

6. Datasets, benchmarks, and evaluation

KITTI is the seminal benchmark but lacks radar and covers a narrow scenario range [5]. nuScenes provides a full camera, LiDAR, and radar suite with 360° coverage and has become a major benchmark for three-modality studies [6]. Waymo Open Dataset is large-scale but has no radar, Argoverse provides no radar. Table 3 summarizes three-modality support [22, 23]. Detection is scored by mAP and, on nuScenes, the composite NuScenes Detection Score (NDS), segmentation is commonly evaluated using IoU, tracking uses AMOTA. Annotation granularity and scenario coverage matter as much as the metric. Few public datasets support all three modalities. Among them, nuScenes remains the most widely used benchmark for camera–LiDAR–radar fusion, although radar-centric datasets are beginning to appear.

Table 3. Representative benchmarks and three-modality support

Dataset	Camera	LiDAR	Radar	3-modal
KITTI [5]	yes	yes	no	no
nuScenes [6]	yes	yes	yes	yes
Waymo [22]	yes	yes	no	no
Argoverse [23]	yes	yes	no	no

7. Challenges and future directions

Several obstacles still bound real-world fusion. Adverse weather and low illumination remain unsolved, particularly for LiDAR [2, 3]. On-vehicle computational constraints limit the amount of cross-modal interaction that can be deployed effectively [1]. Long-range perception and sensor-degradation handling are areas where fusion should provide the greatest benefit but remain relatively under-validated. Simulation-to-real transfer matters because real multi-modal labels are sparse [6].

DOI: 10.54254/2755-2721/2026.LD37798

Radar underuse persists: its sparse and noisy points could often remain a major bottleneck, rather than the fusion operator itself [4].

Three trajectories are likely to dominate near-term work. End-to-end fusion jointly optimizes multiple perception components instead of stacking fixed modules, planning-oriented pipelines such as UniAD exemplify this direction, although UniAD itself is camera-centric rather than a multi-sensor fusion example [24]. Foundation-model perception is emerging as a direction in which multimodal large models interpret driving video and language—DriveGPT4 is a representative example—and may enable more general representations with improved transfer across tasks and modalities [25]. Online, self-supervised calibration may further improve fusion robustness when sensors drift or degrade [8].

8. Conclusion

This survey reviewed camera–LiDAR–radar fusion for autonomous-driving perception through a three-view taxonomy of processing stage, representation, and sensor pair. Four findings follow. BEV unified representation is the de facto backbone because it removes viewport mismatch. Radar remains the least exploited modality, its sparse point cloud, rather than the fusion algorithm, is the main barrier. The practical ceiling is increasingly set by calibration and weather, not by architecture, with data-level fusion most sensitive and decision-level fusion most robust. The field is shifting from hand-crafted modules towards end-to-end and foundation-model designs.

Despite rapid progress, significant challenges remain before multi-sensor fusion can be reliably deployed in production autonomous vehicles. Adverse weather, computational constraints, and the scarcity of large-scale, well-annotated multi-modal datasets continue to limit the scope and robustness of current systems. Online calibration, tighter radar integration, and foundation-model-based perception represent the most promising near-term directions for addressing these limitations. As the field matures, the integration of end-to-end learning paradigms with robust sensor fusion is expected to yield perception systems that are not only more accurate but also more dependable under the diverse and demanding conditions of real-world driving. For practitioners, the practical message is that investment in calibration, synchronization, and radar-specific modelling is likely to pay off more than further architectural complexity.

The review is limited to 2019–2025 English-language venues, so industrial systems and non-mainstream datasets may be under-represented, and the benchmark comparison is a snapshot. Closing that gap, especially for post-2025 foundation-model work, is the most direct next step. Together with online calibration and tighter radar integration, these directions point toward fusion that is not only more accurate, but more dependable under real driving conditions.

References

[1] Wang, Z., Wu, Y., & Niu, Q. (2020). Multi-sensor fusion in automated driving: A survey. *IEEE Access, 8*, 2847–2868. https://doi.org/10.1109/ACCESS.2019.2962554

[2] Roriz, R., Cabral, J., & Gomes, T. (2022). Automotive LiDAR technology: A survey. *IEEE Trans. Intelligent Transportation Systems, 23*(7), 6282–6297. https://doi.org/10.1109/TITS.2021.3086804

[3] Hecht, J. (2018). Lidar for self-driving cars. *Optics and Photonics News, 29*(1), 26–33. https://doi.org/10.1364/OPN.29.1.000026

[4] Giuffrida, L., Masera, G., & Martina, M. (2023). A survey of automotive radar and lidar signal processing and architectures. *Chips, 2*(4), 243–261. https://doi.org/10.3390/chips2040015

[5] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. *IEEE CVPR*. https://doi.org/10.1109/CVPR.2012.6248074

DOI: 10.54254/2755-2721/2026.LD37798

[6] Caesar, H., et al. (2020). nuScenes: A multimodal dataset for autonomous driving. *IEEE CVPR*. https://doi.org/10.1109/CVPR42600.2020.01164
[7] Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., & Heide, F. (2020). Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. *IEEE CVPR*. https://doi.org/10.1109/CVPR42600.2020.01170
[8] Zhu, J., Li, H., & Zhang, T. (2024). Camera, LiDAR, and IMU based multi-sensor fusion SLAM: A survey. *Tsinghua Science and Technology, 29*(2), 415–429. https://doi.org/10.26599/TST.2023.9010010
[9] Vora, S., Lang, A., Helou, B., & Beijbom, O. (2020). PointPainting: Sequential fusion for 3D object detection. *IEEE CVPR*. https://doi.org/10.1109/CVPR42600.2020.00466
[10] Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3D object detection network for autonomous driving. *IEEE CVPR*. https://doi.org/10.1109/CVPR.2017.691
[11] Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3D object detection. *IEEE CVPR*. https://doi.org/10.1109/CVPR.2019.00752
[12] Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q. V., Yuille, A., & Tan, M. (2022). DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection. *IEEE CVPR*. https://doi.org/10.1109/CVPR52688.2022.01667
[13] Pang, S., Morris, D., & Radha, H. (2020). CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. *IEEE IROS*. https://doi.org/10.1109/IROS45743.2020.9341791
[14] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., & Dai, J. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *ECCV*. https://doi.org/10.1007/978-3-031-20077-9_1
[15] Philion, J., & Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. *ECCV*. https://doi.org/10.1007/978-3-030-58568-6_12
[16] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., & Han, S. (2023). BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *IEEE ICRA*. https://doi.org/10.1109/ICRA48891.2023.10160968
[17] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., & Tai, C.-L. (2022). TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. *IEEE CVPR*. https://doi.org/10.1109/CVPR52688.2022.00116
[18] Chen, X., Zhang, T., Wang, Y., Wang, Y., & Zhao, H. (2023). FUTR3D: A unified sensor fusion framework for 3D detection. *IEEE CVPR*. https://doi.org/10.1109/CVPRW59228.2023.00022
[19] Nabati, R., & Qi, H. (2021). CenterFusion: Center-based radar and camera fusion for 3D object detection. *IEEE WACV*. https://doi.org/10.1109/WACV48630.2021.00157
[20] Yang, B., Guo, R., Liang, M., Casas, S., & Urtasun, R. (2020). RadarNet: Exploiting radar for robust perception of dynamic objects. *ECCV*. https://doi.org/10.1007/978-3-030-58523-5_29
[21] Song, J., Zhao, L., & Skinner, K. A. (2024). LiRaFusion: Deep adaptive LiDAR-Radar fusion for 3D object detection. *IEEE ICRA*. https://doi.org/10.1109/ICRA57147.2024.10611436
[22] Sun, P., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. *IEEE CVPR*. https://doi.org/10.1109/CVPR42600.2020.00252
[23] Chang, M.-F., et al. (2019). Argoverse: 3D tracking and forecasting with rich maps. *IEEE CVPR*. https://doi.org/10.1109/CVPR.2019.00895
[24] Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., & Li, H. (2023). Planning-oriented autonomous driving. *IEEE CVPR*. https://doi.org/10.1109/CVPR52729.2023.01712
[25] Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K. K., Li, Z., & Zhao, H. (2024). DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*. https://doi.org/10.1109/LRA.2024.3440097