electronics

MDPI

Review

A Depth-Centric Review of LiDAR–Camera Fusion: From 2D Sensors to High-End Perception for Autonomous Systems

Andrea D'Antona *, Saverio Farsoni, Jacopo Rizzi and Marcello Bonfè

Department of Engineering, University of Ferrara, 44122 Ferrara, Italy; saverio.farsoni@unife.it (S.F.); jacopo.rizzi@unife.it (J.R.); marcello.bonfe@unife.it (M.B.)
* Correspondence: andrea.dantona@unife.it

Abstract

Existing surveys on LiDAR–camera fusion share a blind spot: they assume multi-beam 3D LiDAR as the default sensor, overlooking the wide range of real-world configurations, from a single-plane 2D LiDAR on an indoor robot to a multi-sensor suite on a self-driving car. This review argues that the optimal fusion strategy is determined not by the highest-end architecture, but by the depth information actually available. We introduce the concept of a *depth continuum* and a three-axis taxonomy that maps any LiDAR–camera configuration (defined by LiDAR beam count and camera depth capability) to its appropriate fusion strategy: projection-based, depth completion, Bird's Eye View (BEV), attention-based, or late fusion. Through a structured analysis of the 2021–2026 literature across autonomous driving, industrial AGVs, agricultural robotics, and indoor mobile robots, we derive five key findings: (i) the fusion strategy should track the available depth, with projection and depth completion for sparse LiDAR, BEV for moderate density, and attention-based fusion for dense configurations; (ii) depth completion is the bridging mechanism that supplies low-cost setups with dense depth, but only when trained at the target LiDAR density; (iii) BEV fusion is validated in the literature only from about 32 beams upward, a practical lower bound rather than a controlled threshold; (iv) no standardized benchmark exists for 2D LiDAR–camera fusion, creating a gap between academic research and practical robotics; and (v) late fusion, often dismissed as simplistic, keeps operating when one sensor is lost, a robustness that follows from its architecture but has not yet been validated on large-scale benchmarks. This review aims to serve both as a structured reference for researchers and as a practical configuration-to-strategy mapping for engineers designing perception systems across the full spectrum of autonomous applications.

Keywords: sensor fusion; LiDAR–camera fusion; autonomous vehicles; mobile robotics; depth completion; Bird's Eye View (BEV); 2D LiDAR; 3D LiDAR; SLAM; mapping; object detection; calibration

Check for updates

Academic Editor: Toshifumi Moriyama

Received: 22 September 2026
Revised: 19 October 2026
Accepted: 22 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The ability to perceive and understand the three-dimensional environment is a fundamental requirement for any autonomous system, whether operating on public roads, on factory floors, or in agricultural fields. This perception task demands both precise geometric understanding of the scene and rich semantic interpretation of objects, lanes, obstacles, and free space. No single sensor modality can reliably provide both: Light Detection and Ranging (LiDAR) sensors deliver accurate 3D measurements but produce sparse, colorless

point clouds with no semantic context, while cameras capture dense, textured images rich in appearance information but lack native depth perception.

The complementary nature of these two modalities has made LiDAR–camera fusion the most actively researched approach to robust environmental perception. The past five years have seen a rapid growth of fusion architectures, driven by advances in deep learning, the availability of large-scale multi-modal datasets, and the increasing deployment of autonomous systems in real-world conditions. Several comprehensive surveys have mapped this rapidly evolving landscape [1–15].

However, these reviews share a common limitation: they almost exclusively assume 3D multi-beam LiDAR as the standard sensor configuration, typically 32 beams or more, as found in the KITTI [16], nuScenes [17], and Waymo [18] benchmark datasets. This implicit assumption creates a significant gap between the academic literature and the broader landscape of autonomous systems, where sensor configurations span a wide range of cost, complexity, and depth-sensing capability. At one end of this spectrum, a low-cost indoor mobile robot may operate with a 2D LiDAR and an RGB-D depth camera; at the other end, a fully self-driving car (SAE Level 4 automation) may deploy multiple 128-beam LiDARs, surround-view cameras, and radar. Sensor-centric surveys [19,20] have cataloged this hardware diversity. The optimal fusion strategy depends critically on where a system falls along this spectrum. Organizing the literature around this spectrum, rather than around network architectures as existing surveys do, has three practical advantages. First, it makes the classification predictive: given a sensor configuration, it indicates which fusion strategy is appropriate, instead of only describing existing methods. Second, it covers low-cost configurations, including 2D LiDAR, that architecture-centric surveys omit. Third, it ties each conclusion to the depth regime in which it holds and to the strength of the evidence behind it (Section 13). The relationship between the proposed taxonomy and existing classifications is discussed in Sections 1 and 4.3.

Contributions. This review addresses this gap through the following contributions:

1. **A depth-centric taxonomy** that classifies LiDAR–camera fusion methods along three axes: LiDAR beam configuration (from 2D single-plane to 128+ multi-beam), camera depth capability (monocular, stereo, RGB-D, time-of-flight (ToF), multi-camera surround), and the role of depth information in the fusion strategy, which can be projection-based, depth completion, a shared Bird's Eye View (BEV) representation, depth-aware attention, or depth-independent late fusion.
2. **Full-spectrum coverage**, from low-cost configurations (2D LiDAR + depth camera, typical of indoor robotics and budget-constrained automated guided vehicles (AGVs)) to high-end setups (128-beam LiDAR + multi-camera surround, typical of self-driving cars), analyzing how fusion strategies must adapt to the available depth information.
3. **A cross-domain analysis** comparing fusion approaches across autonomous driving, industrial AGVs, agricultural robotics, and indoor mobile robots, highlighting how application-specific constraints shape fusion architecture choices.
4. **A practical configuration-to-strategy mapping** linking sensor configurations to recommended fusion approaches based on cost, accuracy, and computational constraints, serving as a guide for practitioners designing perception systems.
5. **The identification of open challenges** specific to low-cost and full-spectrum fusion, including the lack of 2D LiDAR benchmarks, the need for depth-completion methods that scale across beam counts, and robustness requirements for safety-critical deployment.

Paper organization. The remainder of this paper is structured as follows. Section 2 describes the review methodology. Section 3 provides an overview of LiDAR and camera depth perception principles and introduces the depth continuum concept. Section 4 presents the proposed depth-centric taxonomy, the core contribution of this review. Section 5 focuses on low-cost fusion with 2D LiDAR and depth cameras. Section 6 compares fusion requirements across the four application domains. Section 7 covers mid-to-high-end 3D LiDAR fusion, including BEV-based approaches. Section 8 examines depth completion as the bridging mechanism across the depth spectrum. Section 9 discusses calibration and synchronization. Section 10 analyzes robustness to adverse conditions and sensor failures. Section 11 surveys available datasets and benchmarks. Section 12 outlines open challenges and future directions, Section 13 summarizes the key findings and the strength of the supporting evidence, and Section 14 concludes the paper.

Comparison with Existing Surveys

Several of the surveys cited above also cover BEV fusion, depth completion, calibration, and robustness; here, we compare how they classify fusion methods, which sensors and domains they consider, and how they treat LiDAR density.

The surveys fall into three groups. The first and largest classifies methods by the stage at which the modalities are combined (early, middle, and late fusion), sometimes adding a class for BEV or transformer-based methods [5,6,8,10,11]. The second proposes finer schemes for LiDAR–camera detection, based on fusion strength [9], on stage, input, and granularity [4], on representation, alignment, and fusion [21], on the granularity of the fused features [1,12], or on the fusion location [2]. The third organizes the literature by task, modality, or system component rather than by fusion scheme [3,7,13–15,22].

All these schemes describe how a network is built, not which sensors it needs. All the surveys address autonomous driving, and none of them discusses 2D LiDAR. LiDAR density appears only indirectly: Wang et al. [4] explain the larger gain of fusion on nuScenes than on KITTI and Waymo with its sparser 32-beam LiDAR, and Qian et al. [5] cite limited beam density as the reason why depth completion is needed, but neither uses density to choose a fusion strategy. Qian et al. also mention agricultural and robotic applications, but only as examples outside their core comparison. Calibration, robustness, and datasets, on the other hand, are already well covered [5,8,11,13].

Against this background, our contribution is not a new list of fusion methods but a different index for them. The three-axis taxonomy keeps the existing labels (Section 4.3 shows how they map onto the five strategies) and adds the LiDAR tier and camera depth capability for which each method works, so that the literature can be read starting from a sensor pair. The depth continuum places all configurations, from a 2D LiDAR with an RGB-D camera to multi-LiDAR rigs, on one scale, which also explains the role of depth completion as a bridge between tiers rather than a separate task. Finally, AGVs, agricultural robots, and indoor robots are part of the corpus rather than examples, and calibration, robustness, and datasets are discussed in relation to the depth regime in which they matter.

2. Review Methodology

The literature search was conducted between January and July 2026 across Scopus, IEEE Xplore, Web of Science, arXiv, and Google Scholar, targeting publications from 2021 onward that address LiDAR–camera fusion for perception tasks. After removing duplicates and screening by title and abstract, 288 publications underwent full-text review, from which a corpus of 111 publications was selected based on relevance to fusion architecture design, depth completion, calibration, robustness, BEV representation, or 2D/3D

LiDAR comparison, applying the search strings, eligibility criteria, and quality assessment detailed below.

Search strings and eligibility criteria. The query was built from four groups of keywords (A–D in Table 1), describing the LiDAR, the camera, the type of integration, and the perception task. A publication was retrieved if its title, abstract, or keywords contained at least one term from each group. Since driving papers dominate the results of this query, we ran it a second time with a fifth group of domain terms (E), so as to also retrieve work on AGVs, agriculture, and indoor robots. The syntax was adapted to each database; Google Scholar, which does not support this kind of query, was used only to follow the references and citations of the included papers. A publication was included if it satisfied all inclusion criteria (I1–I4) and none of the exclusion criteria (E1–E5) listed in Table 1.

Table 1. Search protocol: concept blocks of the Boolean query and eligibility criteria.

ID	Item	Definition
Query blocks (terms within a block joined by OR; blocks joined by AND)		
A	Sensor 1	LiDAR; laser scanner; laser range finder; 2D LiDAR; 3D LiDAR
B	Sensor 2	camera; RGB; RGB-D; stereo; monocular; image
C	Integration	fusion; depth completion; extrinsic calibration; Bird's Eye View; BEV
D	Task	object detection; segmentation; perception; obstacle detection; SLAM; mapping
E	Domain (additional run)	mobile robot; AGV; warehouse; agriculture; agricultural; orchard; indoor
Inclusion criteria (all required)		
I1	Period, language	Published between January 2021 and July 2026, in English
I2	Publication type	Peer-reviewed journal or conference paper, or arXiv preprint later accepted at a peer-reviewed venue or providing public code
I3	Sensors	LiDAR (2D or 3D, any beam count) combined with at least one camera (monocular, stereo, RGB-D), or a step on which such fusion depends (extrinsic calibration, temporal synchronization, depth completion from LiDAR)
I4	Task, evaluation	Perception task (detection, segmentation, depth estimation, mapping/localization, obstacle detection) with quantitative evaluation on a public benchmark or a documented setup
Exclusion criteria (any one sufficient)		
E1	Modality	Single modality, or LiDAR/camera fused only with other sensors (e.g., radar, IMU) without a LiDAR–camera pairing
E2	Scope	Sensor hardware, control, or planning without a perception component
E3	Format	Extended abstract, thesis, patent, or editorial
E4	Reporting	Sensor configuration (LiDAR type and beam count, camera type) not reported, so the work cannot be placed on the depth continuum
E5	Duplication	Superseded by a later, extended version of the same work (only the latter retained)

Quality assessment. Each of the 288 full-text papers was scored on four binary criteria: (Q1) the fusion method is described in enough detail to identify the fusion stage and the role of depth; (Q2) the sensor configuration is fully specified; (Q3) the evaluation uses a public benchmark or a reproducible, documented setup; (Q4) the results are compared against at least one baseline (single-modality or alternative fusion method). Papers satisfying fewer than two criteria were excluded. Screening and quality assessment were performed by the first author; uncertain cases were resolved by discussion with the co-authors.

Strength of evidence. Because the corpus combines results obtained with different networks, datasets, and protocols, the review distinguishes three levels of support for its statements: *measured*, when a result comes from a controlled comparison (same data and protocol, one factor varied) or a dedicated benchmark; *consistently observed*, when a pattern recurs across independent studies that are not directly comparable; and *hypothesized*, when a statement rests on architectural reasoning without direct experimental validation. The key findings and their level of support are summarized in Section 13.

The selected publications were categorized according to the three-axis taxonomy proposed in Section 4. Table 2 summarizes their distribution across fusion strategies and application domains. A small number of foundational works pre-dating the search window (the KITTI, nuScenes, and Waymo benchmark datasets, and seminal baseline architectures such as PointPillars, PointPainting, Sparse-to-Dense, and CSPN) are also cited throughout the review as standard reference points against which the 2021–2026 corpus is compared; they fall outside the systematic search and are not counted among the 111 corpus publications classified in Table 2. The same applies to the robotics datasets and evaluation references used in Section 11 (TorWIC, Sugar Beets 2016, Robot@Home, OpenLORIS-Scene, JRDB, the MIT Stata Center dataset, a recent intralogistics dataset, and Semantic2D), which were retrieved specifically for that analysis.

Effect of the time window. We start the window in 2021 because the earlier literature is already covered by previous surveys [8,9,22], and because the following years are the ones in which BEV and transformer-based fusion appeared and fusion started to be applied outside automotive benchmarks. This choice has a cost. The counts in Table 2 describe where research effort went in 2021–2026, not the overall weight of each strategy: projection-based fusion and depth completion, already established before 2021, enter the corpus only through their recent extensions, whereas BEV fusion, whose LiDAR–camera variants appeared from 2022, is counted from its first papers. The table is therefore a picture of current activity, not a ranking of strategies. To avoid a narrative built on works that were not screened with the same criteria, we use the pre-2021 works only as baselines and historical reference points; the evidence behind the findings summarized in Section 13 comes from the corpus, except for finding (iv), which rests on the review of available datasets (Section 11).

Table 2. Distribution of the reviewed corpus (111 publications) by primary fusion strategy and application domain. Domains: Driving = autonomous driving; AGV = industrial AGVs and logistics; Agri. = agricultural robotics; Indoor = indoor mobile robots.

Fusion Strategy	Driving	AGV	Agri.	Indoor	Total
Projection-based	8	4	2	6	20
Depth completion	10	4	2	2	18
BEV fusion	17	2	1	0	20
Attention-based	9	1	1	1	12
Late fusion	5	4	2	3	14
Calibration	9	1	0	1	11
Robustness/Adverse cond.	8	0	1	0	9
Datasets/Benchmarks	4	1	2	0	7
Total	70	17	11	13	111

Three patterns in Table 2 are relevant to the rest of the review. First, autonomous driving accounts for 70 of the 111 publications, and the strategies that need dense depth are concentrated there: 17 of the 20 BEV papers and 9 of the 12 attention-based papers target driving, and no BEV paper targets indoor robots. Second, the domains built around 2D or low-beam LiDAR rely on strategies that need little depth: projection-based and late fusion account for 9 of the 13 indoor papers and 8 of the 17 AGV papers. Third, 12 of the 16 robustness and dataset papers concern driving and none concerns indoor robots, which is the benchmark gap discussed in Section 11.2. These patterns agree with the depth continuum, but they do not prove it: domain and sensor configuration are tied to the available datasets, and driving research adopts BEV fusion partly because nuScenes and Waymo provide surround cameras and 32+ beam LiDAR. For this reason, finding (i) is classified as consistently observed rather than measured (Section 13).

3. Fundamentals of Depth Perception in Robotics

This section establishes the conceptual foundation for understanding how LiDAR and camera sensors contribute depth information and why the quantity and quality of this information fundamentally determine the appropriate fusion strategy.

3.1. LiDAR: Principles and Depth Characteristics

LiDAR sensors operate on the time-of-flight principle, emitting laser pulses and measuring the round-trip time to determine distance with centimeter-level accuracy. The key parameter differentiating LiDAR configurations is the number of laser beams (or channels), which directly determines the density and coverage of the resulting 3D point cloud.

2D LiDAR. A 2D LiDAR (also called single-plane or planar LiDAR) emits a single laser beam that rotates around one axis, producing distance measurements in a single horizontal plane. Typical models such as the RPLidar A1/A2/A3, YDLidar X4, or Hokuyo URG series provide 360° coverage with ranges of 10–30 m at update rates of 5–15 Hz, at costs ranging from $100 to $500 [23]. These sensors are ubiquitous in indoor mobile robotics for simultaneous localization and mapping (SLAM) and navigation. The critical limitation is the complete absence of vertical information, since a 2D LiDAR cannot determine whether an object is a wall, a table, or a low obstacle at different heights.

3D LiDAR. A 3D LiDAR employs multiple laser beams arranged vertically, producing a three-dimensional point cloud. The beam count broadly defines four tiers:

- **Low-beam (16–32 channels):** Velodyne VLP-16, Hesai PandarXT-32, Livox Mid-70. These provide sparse 3D coverage at moderate cost ($2000–$4000); compact non-repetitive scanning units such as the Unitree L2 now cost a few hundred dollars, at the price of a shorter range (about 30 m). A 16-beam LiDAR generates approximately 300,000 points per second, sufficient for basic 3D perception but requiring significant interpolation for dense scene understanding.
- **Mid-beam (64 channels):** Velodyne HDL-64E, the sensor used in the KITTI benchmark. It generates approximately 1.3 million points per second and has become the *de facto* standard in autonomous driving research.
- **High-beam (128+ channels):** Ouster OS1-128, RoboSense RS-Ruby Plus, Hesai AT128. These produce 2.5–5 million points per second, approaching quasi-dense depth.
- **Solid-state and emerging technologies:** micro-electro-mechanical system (MEMS)-based, flash, and optical phased-array LiDARs offer alternative scanning patterns and form factors, with beam-equivalent specifications varying widely [24].

The cost–accuracy trade-off. Table 3 summarizes the key characteristics across LiDAR configurations. The progression from 2D to 128-beam 3D spans roughly two orders of magnitude in cost (three once multi-LiDAR suites are included) and a comparable spread in point-cloud density, a difference that fundamentally changes which fusion strategies are viable. The wide cost range for 128+ beam LiDAR reflects the rapidly evolving market: solid-state units (e.g., Hesai AT128, RoboSense RS-LiDAR-M1) start near $1000–$4000 in automotive volume (which is why a high-beam solid-state device can undercut an older low-beam mechanical one), while traditional mechanical/spinning units (e.g., Ouster OS1-128, Velodyne Alpha Prime) remain at $10,000–$25,000. Range figures in the table are maximum values at high target reflectivity and overlap across tiers because newer low-beam sensors can outrange older 64-beam units. Prices are indicative and subject to volume discounts and regional variation.

Table 3. Comparison of LiDAR configurations and their depth characteristics.

Type	Channels	Points/s	Range (m)	Cost (USD)	Typical Use
2D planar	1	N/A (plane)	10–30	$100–$500	Indoor robots, SLAM
3D low-beam	16–32	∼300 K	80–120	$0.4K–$5K	Drones, AGVs, budget vehicles
3D mid-beam	64	∼1.3 M	100–120	$4K–$12K	Driving research (KITTI standard)
3D high-beam	128+	∼2.5–5 M	120–250	$4K–$25K	Self-driving cars, robotaxis
Multi-LiDAR	64+×N	up to 15 M	varies	$40K+	Full autonomy, mapping

3.2. Cameras: From RGB to Depth

Camera sensors provide dense, textured images but lack native depth information. The degree to which a camera setup can recover depth largely determines its role in the fusion process.

Monocular cameras offer the simplest and cheapest configuration but provide no native depth. Depth must be estimated through monocular depth estimation (MDE) networks, a challenging and ill-posed problem that has progressed considerably with transformer-based architectures and foundation models [25], yet remains fundamentally ambiguous at scale. Without an external metric anchor (a known object size, a fixed camera height, or a second sensor), the network's output is a relative, self-consistent estimate rather than a measurement, and it degrades silently on inputs unlike its training distribution rather than failing detectably. A single camera also offers no physical redundancy: any degradation of that one channel (glare, occlusion, darkness) removes the depth signal entirely, rather than merely reducing its density or accuracy. Monocular cameras are the primary visual sensor in autonomous driving datasets (nuScenes uses six surround monocular cameras), but their lack of depth makes fusion with LiDAR essential rather than optional.

Stereo cameras estimate depth from binocular disparity, providing dense but range-limited depth (up to 50–100 m depending on baseline and resolution). The depth quality degrades quadratically with distance and is sensitive to texture-poor regions and lighting conditions. Stereo cameras represent a middle ground: they provide native depth, albeit less accurate than LiDAR, and their depth estimates can serve as a bridge between sparse LiDAR measurements and dense image appearance [26,27].

RGB-D (color-plus-depth) and depth cameras use structured light (e.g., Intel RealSense D400 series) or indirect ToF (e.g., Microsoft Azure Kinect) to produce dense depth maps at close range (0.3–10 m). These sensors are widely used in indoor robotics and provide the richest depth information per dollar, but are limited to indoor or short-range outdoor use due to ambient light interference [23,28].

Why pair a LiDAR with an RGB-D camera? An RGB-D sensor already measures depth in hardware, so pairing it with a 2D LiDAR (Section 5) may seem redundant, but four gaps remain. Range: RGB-D depth is reliable only to 5–10 m, short of the 10–30 m needed to plan ahead. Field of view: ∼70–90° against the LiDAR's full 360°. Illumination: structured-light and ToF depth degrade under direct sunlight. Localization: RGB-D-only odometry drifts over long trajectories, whereas 2D LiDAR yields a metrically consistent map. RTAB-Map-based systems [23,28,29] fuse the two for exactly this reason: LiDAR for global consistency and 360° coverage, RGB-D for short-range dense depth and visual features.

Multi-camera surround systems (typically 4–6 cameras covering 360°) are the standard configuration for autonomous driving, providing wide coverage without native depth. The challenge of fusing surround-view images with LiDAR in a unified representation has motivated the development of BEV architectures [30,31].

Why not just add more cameras? Cameras are cheap, and overlapping views can themselves yield a point cloud via multi-view stereo, so could a large enough array replace

LiDAR? The limitation is not viewpoint count but what triangulation recovers: multi-camera depth is inferred, not measured. It stays relative rather than absolute, since scale needs an external anchor (a known baseline, an IMU, or a LiDAR point cloud), and accuracy degrades quadratically with distance. More cameras also multiply calibration and synchronization cost (Section 9) without fixing passive sensing's core weakness: in low light or glare, every camera degrades at once, while LiDAR, being active, is largely unaffected (Section 10). This is why, despite progress in vision-centric BEV perception [32], camera-only stacks remain the exception in the surveyed literature.

3.3. Why Sensor Fusion? Complementarity of Depth and Appearance

Before introducing the depth continuum, it is worth recalling why LiDAR–camera fusion is needed in the first place. The motivation lies in the complementarity of the two modalities across three dimensions:

- **Geometric accuracy:** LiDAR provides 3D coordinates with centimeter-level accuracy, while camera-based depth is either relative (monocular estimation, affected by scale ambiguity) or increasingly inaccurate at long range (stereo). Fusing LiDAR anchor points with camera appearance can recover dense depth with geometric precision [33,34].
- **Semantic richness:** Cameras capture texture, color, and fine-grained appearance patterns that enable object classification, lane detection, and traffic sign recognition, tasks that are extremely challenging from sparse, colorless point clouds alone. LiDAR provides the spatial scaffolding onto which semantic labels can be projected [35,36].
- **Environmental robustness:** LiDAR is active (provides its own illumination) and largely invariant to ambient lighting, but degrades in rain, fog, and snow due to backscattering and attenuation. Cameras are passive and struggle in low-light conditions, but are less affected than LiDAR by backscattering in precipitation. The fusion of both modalities can provide graceful degradation rather than catastrophic failure when one sensor is compromised [37–39].

These complementary properties (geometry from LiDAR, semantics from cameras, and resilience from their combined failure modes) motivate the full spectrum of fusion approaches discussed throughout this review.

3.4. The Depth Continuum: From Sparse to Dense

A central conceptual contribution of this review is the depth continuum, the observation that LiDAR–camera configurations can be arranged along a continuous axis of depth information density and quality, as illustrated in Figure 1.

The position of a system along this continuum has direct implications for fusion architecture design (the role of depth in each strategy is detailed in Section 4):

- At the sparse end (2D LiDAR, 16-beam), the depth information is too scarce to support volumetric fusion. Instead, depth serves primarily as a projection tool to associate LiDAR points with image regions, and depth completion becomes essential to densify the sparse measurements.
- At the middle of the continuum (32–64 beams), depth is sufficient to build volumetric representations (voxels, pillars, BEV grids) that can be fused with image features projected into the same space, which is the BEV fusion paradigm.
- At the dense end (128+ beams, including multi-LiDAR setups), the LiDAR point cloud is dense enough that depth completion becomes less critical, and fusion shifts toward attention-based mechanisms that use the dense geometry to guide where and how to extract image features.



Figure 1. The depth continuum: LiDAR–camera configurations arranged by depth density (sparse → dense) and depth accuracy. Each marker represents a sensor configuration; marker shape indicates camera type, color indicates LiDAR tier. The background zones map regions of the continuum to the recommended fusion strategy: projection-based and depth completion for sparse configurations (blue), BEV fusion for moderate depth (orange), and attention-based fusion for dense configurations (red).

4. A Depth-Centric Taxonomy of LiDAR–Camera Fusion

This section presents the core contribution of this review: a three-axis taxonomy for classifying LiDAR–camera fusion methods based on the sensor configuration's depth characteristics and the role that depth plays in the fusion process. Figure 2 provides a visual overview of the taxonomy, showing how the three axes (LiDAR depth capability, camera depth capability, and fusion strategy) intersect to define the space of possible configurations.

The five strategies are not new in themselves: they correspond to the early (data-level), middle (feature-level), and late (decision-level) fusion classes of previous surveys (Section 4.3). What the taxonomy adds is that it indexes them by the depth the sensor pair provides, including 2D LiDAR configurations (Section 1). Calibration is not one of the strategies but a prerequisite for all of them: CalibRefine [40], for instance, also combines projection and attention, but to estimate the alignment between the sensors rather than to fuse their data (Section 9).

The figure reads as a lookup table: its columns are the LiDAR tiers (2D to 128+ beams), its rows are the camera types (monocular to ToF), and each cell indicates the strategy recommended for that pair. Cells marked "N/A" are combinations of little practical use: RGB-D and ToF cameras work only at short range (<10 m), so pairing them with long-range 64+ beam LiDAR is meaningless outside niche multi-range rigs; multi-camera surround setups presuppose 3D depth for view stitching and are seldom paired with a single-plane 2D LiDAR, though low-cost AGVs occasionally do. The "16–32 beam" column straddles

the transition point of the depth continuum: at 16 beams, the cloud is very sparse for voxel-based processing, and the surveyed low-beam systems rely on projection and depth completion, whereas 32 beams (the nuScenes configuration) is where most BEV methods are trained and validated. This boundary is an indicative one, derived from where BEV fusion has been demonstrated rather than from controlled comparisons (Section 7.1); the arrow in those cells ("DC → BEV", with DC short for depth completion) marks this shift. The two sensor axes and the fusion strategies are detailed next.

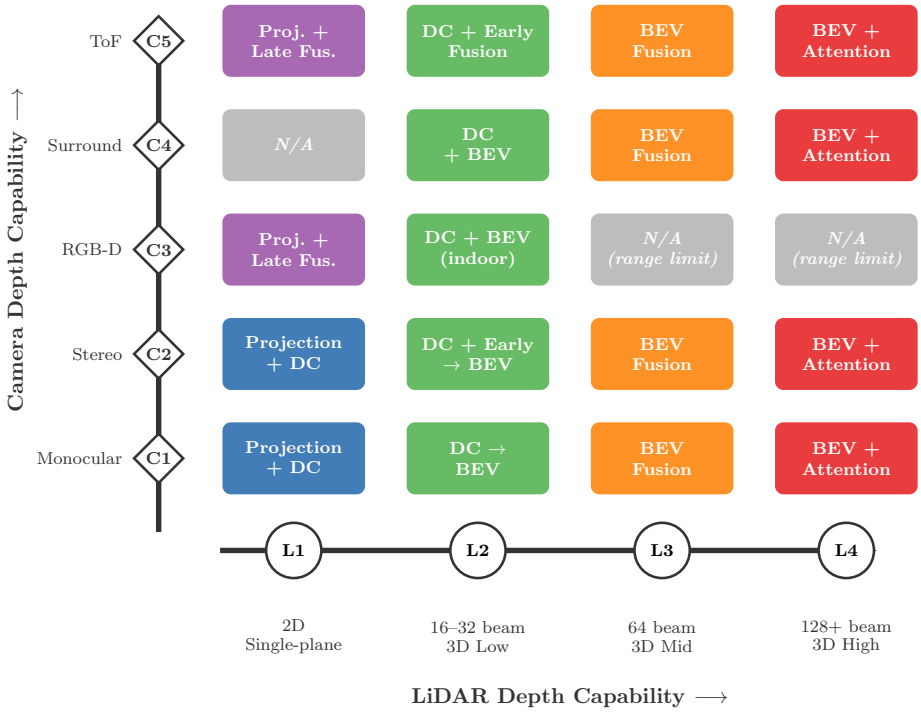

Figure 2. Depth-centric taxonomy as a configuration matrix. Columns: LiDAR depth capability (L1–L4). Rows: camera depth capability (C1–C5). Each cell is the fusion strategy recommended for that pair; an arrow ("DC → BEV") marks a transitional cell and "N/A" an impractical pairing (both explained in the text). Colors: ■ projection-based, ■ depth completion, ■ BEV fusion, ■ attention-based, ■ late fusion, ■ not applicable (N/A).

4.1. The Two Sensor Axes

LiDAR depth capability ranks the LiDAR by how much of the 3D scene it samples: **L1**, 2D single-plane LiDAR (one plane, no vertical resolution; indoor robots, low-cost AGVs, education); **L2**, 3D low-beam, 16–32 beams (sparse 3D; drones, budget vehicles, outdoor robots); **L3**, 3D mid-beam, 64 beams (semi-dense 3D; the autonomous-driving research standard, used in KITTI); **L4**, 3D high-beam, 128+ beams (dense 3D; robotaxis). Multi-LiDAR rigs such as Waymo's five-LiDAR setup belong to L4, combining several such sensors for near-complete 360° reconstruction.

Camera depth capability ranks the camera by how well it recovers depth unaided: **C1**, monocular RGB (no native depth, needs learned estimation; the norm in nuScenes and Waymo); **C2**, stereo (depth from disparity, range-limited and range-dependent in accuracy); **C3**, RGB-D/depth camera (dense depth at short range via structured light or ToF, mostly indoor); **C4**, multi-camera surround (360° coverage, no native depth; standard on driving platforms); **C5**, ToF camera (dense depth at medium range via direct per-pixel time-of-flight, emerging in automotive use).

4.2. The Role of Depth in Fusion

The third axis is the most novel: it classifies fusion strategies by the role that depth information plays, which changes with how much depth is available.

Depth as a projection tool (data-level fusion). When depth is extremely sparse (2D or 16-beam LiDAR), its only job is to establish geometric correspondence: LiDAR points are projected onto the image plane through the extrinsic calibration, and each point inherits the semantics of the pixel it lands on. PointPainting [41] appends 2D segmentation scores to each point; PointAugmenting [42] appends convolutional neural network (CNN) features instead [22,43,44]. The fusion is one-directional (camera → LiDAR) and reaches only points inside the camera frustum, but it needs no learned depth and is computationally cheap.

Depth completion (sparse-to-dense bridge). For sparse LiDAR (2D to 32 beams), depth completion turns the sparse but accurate anchor points into a quasi-dense depth map under image guidance, effectively upgrading the system's depth capability by inference [25,45,46]. MSMDFusion [47] associates points with image features at multiple depth-seed levels; DMFusion [48] merges LiDAR and stereo-derived depth into one coherent map; FastFusion [27] injects LiDAR priors into every stage of stereo matching for real-time dense output.

Depth as a shared representation (BEV fusion). When LiDAR depth is moderately dense (32–64 beams), both modalities can be lifted into a common Bird's Eye View grid and fused there [31,49]. LiDAR projects trivially since its points are already in 3D; camera features must be lifted using inferred depth. BEVFusion [50] uses a learned depth distribution for the lift, while BEVFormer [30] (though camera-only) contributed the spatio-temporal transformer backbone that later LiDAR–camera BEV methods reuse.

Depth-aware attention fusion. For high-beam LiDAR (64–128+ beams), the dense geometry is reliable enough to bias attention: it decides where to look while the camera decides what is there. GAFusion [51] lets LiDAR guide the attention weights over camera features; DeepInteraction [52] keeps the two representations separate and couples them through multi-input, multi-output cross-attention rather than collapsing them into one feature. Related architectures [53–56] and efficient variants for service robots [57] apply the same principle.

Depth-independent late fusion. At any depth level, late (decision-level) fusion runs each modality's detector independently and combines the output boxes [58,59]. It forgoes cross-modal feature interaction and usually trails middle fusion when depth is available, but it is sensor-agnostic and, by construction, keeps producing detections when one sensor is lost (a property that follows from the architecture rather than from benchmark measurements; Section 10), making it the right choice when sensors are heterogeneous, asynchronous, or when safety-critical redundancy matters most.

4.3. Comparative Analysis

The five strategies map onto the stage-based classes of previous surveys (Section 1): projection-based fusion is early (data-level) fusion, BEV and depth-aware attention are middle (feature-level) fusion, and late fusion is decision-level fusion. Depth completion has no single stage, since the dense depth map it produces can feed any of the others; this is why stage-based surveys treat it as a separate task [5,22]. None of these classes, however, records the depth a method needs, so methods with the same label can suit very different sensors: projection-based decoration works with any LiDAR whose points fall in the image, including 2D LiDAR, where it is the most common strategy (Table 2 and Section 5), whereas BEV methods trained at 32 beams degrade on sparser input (Section 7.1). The depth-centric taxonomy adds this information.

Table 4 summarizes the five strategies on common criteria, drawing on Sections 8–10; the depth regimes are indicative (Section 7.1).

Table 4. Systematic comparison of the five fusion strategies: typical depth regime, main advantages and limitations, and suitable application scenarios.

Strategy	Depth Regime	Advantages	Limitations	Suitable Scenarios
Projection-based	2D, 16-beam	Low computational cost; no learned depth; simple to implement	One-directional (camera to LiDAR); limited to points inside the camera frustum; most sensitive to calibration errors	Indoor robots, low-cost AGVs, embedded platforms
Depth completion	2D to 32-beam	Dense depth from sparse LiDAR; applicable down to 2D LiDAR	Must be trained at the target LiDAR density; errors grow far from LiDAR points; added computation	Budget vehicles, drones, agricultural robots
BEV fusion	32 to 64-beam	Common top-down frame aligned with planning; extensible to more sensors and tasks	Documented from about 32 beams (indicative); camera lift depends on predicted depth; sensitive to extrinsic errors; high computational cost	Autonomous driving, production-oriented stacks
Attention-based	64 to 128+ beam	Learned soft association; strong reported accuracy on driving benchmarks; least sensitive to extrinsic errors	High computational and memory cost; requires dense geometry and large training sets	Robotaxis, multi-LiDAR platforms
Late fusion	Any	Modular and sensor-agnostic; tolerates asynchronous sensors; keeps operating if one sensor is lost (by construction)	No cross-modal feature interaction; lower nominal accuracy; no protection against a faulty but active sensor	Heterogeneous or asynchronous sensors, safety-critical redundancy

Table 5 provides a quantitative comparison of representative methods across the main fusion strategies on the nuScenes benchmark, reported as mean Average Precision (mAP) and the nuScenes Detection Score (NDS) (both defined in Section 11). As a reference point, PointPillars [60] is included as a LiDAR-only baseline: it encodes the point cloud into vertical "pillars" that are processed by a 2D convolutional backbone, without any camera input, and its scores mark the performance achievable from geometry alone before fusion is introduced. All scores are from the nuScenes test server; validation results are not mixed in, since the val–test gap is not systematic when test submissions are trained on train+val (BEVFusion reports 68.5/71.4 mAP/NDS on val and 70.2/72.9 on test [50]). The methods still differ in backbone, input resolution, training schedule, and year (2019–2024), so the table shows a trend rather than a ranking. The robust observation is that every fusion method improves on the LiDAR-only baseline by 16 to 43 mAP points. The gap between PointPainting and the newer methods largely reflects backbone progress since 2019, and the spread among BEV, depth-completion, and attention-based methods (70.2–73.6 mAP) is within the range of implementation choices, so no ranking among them should be

inferred. The table fixes the LiDAR density (nuScenes uses a 32-beam sensor); how fusion performance scales with beam count is examined separately in Section 8. A comparable trend is reported on KITTI, where cross-modal augmentation of the point cloud with image features improves 3D detection over the LiDAR-only baseline [41,42].

Table 5. Performance comparison of representative fusion methods on the nuScenes benchmark, all evaluated on the test server. Values are from the respective original publications. Methods differ in backbone, training schedule, and publication year, so the table indicates a trend rather than a controlled comparison of fusion strategies.

Method	Fusion Strategy	nuScenes mAP/NDS (Test)
PointPillars [60] (LiDAR only)	N/A	30.5/45.3
PointPainting [41]	Projection-based	46.4/58.1
BEVFusion [50]	BEV fusion	70.2/72.9
MSMDFusion [47]	Depth completion	71.5/74.0
GAFusion [51]	Attention-based	73.6/74.9

5. Low-Cost Fusion: 2D LiDAR and Depth Camera Approaches

This section focuses on the sparse end of the depth continuum, where 2D LiDAR and depth cameras (RGB-D, short-range stereo) are the primary sensors. These configurations dominate indoor robotics, educational platforms, budget-constrained AGVs, and agricultural robots.

5.1. The 2D LiDAR + Camera Configuration

The typical low-cost perception stack combines a 2D LiDAR (e.g., RPLidar, YDLidar) providing 360° planar distance measurements with either an RGB-D camera (Intel RealSense, Microsoft Azure Kinect, Orbbec Astra) or a monocular/stereo camera. The total sensor cost ranges from $300 to $2000, making these configurations accessible for a wide range of applications [23,61].

The key perception tasks for these configurations are as follows:

- **SLAM and navigation:** The 2D LiDAR provides accurate planar localization (2D SLAM via GMapping, Cartographer), while the camera enables visual place recognition, loop closure detection, and semantic mapping [62–65].
- **Object detection and tracking:** The camera detects and classifies objects in 2D, and LiDAR distance measurements provide metric scale and obstacle ranging [61,66].
- **Semantic mapping:** RGB-D point clouds projected to 3D and filtered with 2D LiDAR data create enriched maps for navigation and interaction [67].

5.2. Depth Projection and Alignment for 2D LiDAR

With a single-plane LiDAR, projection-based fusion becomes the primary feasible approach. Li et al. [43] proposed a practical system integrating 2D LiDAR with YOLOv9-based camera detection within the Robot Operating System (ROS) 2 for advanced driver-assistance system (ADAS) applications, achieving real-time obstacle detection at low computational cost. The 2D LiDAR provides distance measurements to objects detected by the camera, compensating for the monocular camera's lack of scale.

For SLAM applications, the integration of 2D LiDAR with RGB-D cameras has been extensively studied. Mobile robot navigation systems based on RTAB-Map (Real-Time Appearance-Based Mapping), a widely used graph-based SLAM framework with loop-closure detection, fuse 2D LiDAR scans with RGB-D point clouds to improve both mapping accuracy and localization efficiency [23,28,29]. A common strategy projects RGB-D point clouds onto the 2D plane for denoising alongside LiDAR data, applying late fu-

sion to combine the processed data streams. This approach preserves the real-time performance of 2D SLAM while enriching the map with visual features. Tightly coupled multi-sensor fusion frameworks for SLAM [68,69] and mobile robot SLAM systems [70] further demonstrate that fusing LiDAR with visual and inertial data improves robustness in dynamic environments.

5.3. Depth Completion from 2D LiDAR

The extreme sparsity of 2D LiDAR (a single plane of points) makes depth completion both more essential and more challenging than for 3D LiDAR. Gao et al. [23] demonstrated that projecting RGB-D point clouds to the 2D plane and fusing with LiDAR data can effectively densify the depth information. Recent advances in unsupervised depth completion [71] and foundation model distillation for depth completion [25] offer promising directions for inferring dense 3D structure from minimal LiDAR measurements.

5.4. Limitations of 2D LiDAR and the Case for Upgrading to 3D

2D LiDAR configurations have fundamental limitations that cannot be overcome through algorithmic improvements alone:

1. **No vertical discrimination:** A 2D LiDAR cannot distinguish between objects at different heights within the same planar slice, leading to false positives (detecting overhanging objects as obstacles) and false negatives (missing low obstacles below the scan plane).
2. **No 3D object detection:** Full 3D bounding box estimation is impossible from a single-plane scan without strong assumptions about object shape and position.
3. **Limited outdoor capability:** 2D LiDARs typically have shorter range (10–30 m) and are affected by ambient light in outdoor settings.

The decision to upgrade from 2D to 3D LiDAR should be driven by the specific perception requirements: if the application requires 3D object detection, operation in unstructured outdoor environments, or vertical scene understanding, a 16-beam 3D LiDAR (approximately $2000–$4000) represents the minimum viable upgrade, while compact units such as the Unitree L2 (a few hundred dollars) are a cheaper option for short-range indoor and outdoor robots. Table 6 summarizes the trade-offs due to the LiDAR tier alone; the contribution of the paired camera is common to both configurations.

Table 6. Trade-offs between 2D and 3D LiDAR configurations, independent of the paired camera.

Criterion	2D LiDAR	3D LiDAR
LiDAR cost (USD)	$100–$500	$400–$25,000
3D object detection from LiDAR alone	Not possible	Yes (sparse to dense)
Vertical discrimination from LiDAR alone	No	Yes
Depth completion difficulty	Extreme (single plane)	Moderate (depending on beams)
LiDAR-only SLAM capability	2D only	3D (LOAM, LIO-SAM)
Outdoor range	10–30 m	80–250 m
Fusion strategy	Projection, DC, late fusion	DC, BEV, attention
Typical applications	Indoor robots, education, budget AGVs	Self-driving cars, outdoor robots, drones
Computational load	Low	Medium to high

6. Cross-Domain Analysis of Fusion Requirements

Fusion requirements are shaped as much by the operational context as by the sensor hardware. This section compares how the four application domains covered in this review translate the depth-centric taxonomy into concrete design choices, spanning the full

continuum from 2D-LiDAR indoor robots to multi-LiDAR robotaxis. Figure 3 provides a comparative overview of the four domains across six key requirement dimensions.

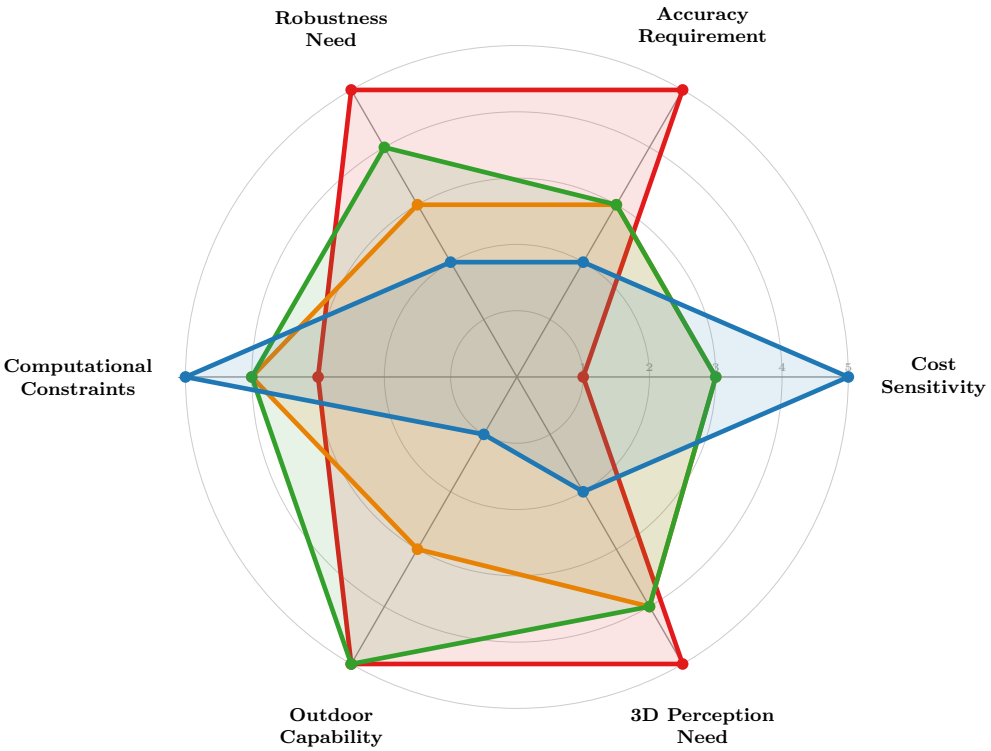

Figure 3. Radar chart comparing the requirement profiles of four application domains for LiDAR–camera fusion. Six dimensions are scored from 1 (low) to 5 (high). The divergent profiles highlight why a single fusion strategy cannot serve all domains. ■ Autonomous driving (L2–L4), ■ industrial AGVs and logistics, ■ agricultural robotics, ■ indoor mobile robots.

Indoor mobile robots. Warehouse AGVs, hospital service robots, and domestic platforms operate in structured environments with flat floors, known layouts, and controlled lighting. Here, 2D LiDAR + RGB-D is the dominant configuration: the LiDAR handles planar localization and obstacle avoidance, while the RGB-D camera provides short-range 3D perception for object manipulation and people detection. The primary fusion strategy is projection-based (Section 5). Computational constraints are tight (embedded ARM or low-power x86), favoring late fusion and lightweight architectures [57].

Agricultural and outdoor robots. Agricultural robots (weeding, harvesting, monitoring) and outdoor logistics platforms face unstructured terrain, varying illumination, dust, and vegetation. These applications increasingly adopt 16–32 beam LiDAR combined with stereo or multi-camera setups for 3D scene understanding [61]. The fusion strategy shifts toward depth completion and early BEV representations to handle sparse 3D data in wide-open spaces. Cost sensitivity remains higher than in autonomous driving but lower than indoor robotics, with typical sensor budgets of $5000–$15,000.

Industrial AGVs and logistics. Automated guided vehicles in factories and ports occupy a middle ground: semi-structured environments where 3D perception adds value (stacking, pallet picking, outdoor–indoor transitions) but budgets are constrained. The trend is toward 16–32 beam LiDAR with monocular or stereo cameras, using depth completion to bridge the gap between sparse LiDAR and dense perception needs [43].

Autonomous driving (L2–L4). At the high end of the depth spectrum, this domain spans driver-assistance passenger cars up to full-autonomy robotaxis, which deploy 64–128+ beam LiDAR (single or multiple units) with surround-view cameras, where BEV and attention-based fusion dominate. In this domain, the sensor budget is the least constrained (typically $15,000–$75,000 for a single-LiDAR perception suite, rising with multi-LiDAR robotaxi rigs), and the primary optimization targets are safety-critical reliability and long-range detection (200+ m). The perception stack must handle diverse object classes (vehicles, pedestrians, cyclists, traffic signs), complex traffic interactions, and operation at highway speeds. BEV fusion has become the *de facto* standard for production-oriented systems because its top-down representation naturally aligns with motion planning and enables multi-sensor extensibility [31]. At the highest end of this domain, full-autonomy platforms such as Waymo (5 LiDARs + 5 cameras) push fusion further toward attention-based architectures where dense geometry from multiple LiDARs guides semantic extraction from surround images [18]: the dense multi-LiDAR configuration enables near-complete geometric coverage, so the design problem is no longer depth scarcity but the efficient exploitation of redundant multi-modal information. The introduction of 4D radar as a third modality, as in RAF (Reliability-Aware Fusion) [72], which learns per-sensor confidence estimates and reweights each modality accordingly, signals a shift toward reliability-aware multi-modal systems. The contrast between this domain and the low-cost configurations discussed above underscores the need for the depth-centric taxonomy proposed in this review: a one-size-fits-all fusion strategy does not exist.

7. Mid-to-High-End Fusion: 3D LiDAR and Multi-Camera Systems

7.1. 3D LiDAR (16–64 Beams) with Mono/Stereo Camera

As depth density increases from 16 to 64 beams, the role of fusion shifts from depth completion toward volumetric representation. At 16 beams, depth completion remains essential, since the point cloud is very sparse for voxel-based processing. At 32 beams, the density of the nuScenes LiDAR, BEV fusion is routinely trained and validated; below it, evidence for BEV fusion is scarce. At 64 beams (the KITTI standard), BEV fusion is the dominant paradigm.

Strength of the evidence for the 32-beam boundary. This boundary should be read as an indicative, empirically observed lower bound, not as a general threshold. It rests on three lines of evidence, none of which is a controlled comparison between fusion strategies: (i) most BEV fusion methods are developed and evaluated on nuScenes, so 32 beams is the lowest density at which their performance is documented, which shows that BEV fusion works there but not that it fails below; (ii) depth completion studies across densities [73] and early- versus mid-level fusion comparisons across beam counts [44] show the trend with density, but use different networks, datasets, and metrics from the BEV literature; and (iii) the only controlled evidence we found is the beam-reduction corruption of the MultiCorrupt benchmark [74], which evaluates the same detectors on nuScenes with the 32 beams reduced to 16, 8, and 4. There, the tested fusion detectors retain on average only about 63–79% of their NDS (68% for BEVFusion), which is consistent with the boundary. However, the models are not retrained at the reduced density and no depth-completion baseline is included, so this result shows that detectors trained at 32 beams degrade on sparser input, not that BEV fusion is inferior to depth completion below 32 beams. Establishing the boundary would require the same scenes captured at multiple LiDAR densities, with each strategy trained and evaluated at each density (Section 12).

A systematic comparison of early- and mid-level fusion across different beam configurations [44] demonstrates that the performance gap between early and mid-level fusion

narrows as beam count increases, which is consistent with the central premise of this review. GAFusion [51], discussed in Section 4, exemplifies the state of the art at this tier.

7.2. BEV Fusion Paradigm

BEV fusion has emerged as the dominant paradigm for 3D perception in autonomous driving, and its rise is directly linked to the availability of sufficiently dense LiDAR point clouds. BEV offers several advantages [31,75]:

1. **Shared representation space:** Both LiDAR (3D points → BEV via projection) and camera (2D features → BEV via depth-based lifting) contribute to the same top-down grid, enabling natural feature-level fusion without explicit 3D-2D correspondence.
2. **Natural integration with planning:** The BEV representation aligns with the ego-vehicle's motion planning space, enabling end-to-end perception-to-planning pipelines.
3. **Occlusion handling:** BEV naturally represents occluded regions as empty grid cells, which can be propagated temporally.
4. **Multi-sensor extensibility:** BEV grids can incorporate features from radar, maps, and temporal history using the same spatial indexing.

Figure 4 illustrates the canonical BEV fusion pipeline, showing the parallel processing streams for LiDAR and camera data and their convergence in the shared BEV representation space. On the LiDAR side, the raw point cloud is first discretized so that a convolutional backbone can process it: *voxelization* partitions 3D space into a regular grid of small cubic cells (voxels) and summarizes the points falling in each, whereas *pillarization* (introduced by PointPillars [60]) replaces the cubes with vertical columns of unlimited height, collapsing the vertical dimension so that the point cloud becomes a 2D pseudo-image that a standard 2D convolutional network can process far more cheaply. Either way, the output is the top-down LiDAR BEV feature map. On the camera side, per-pixel image features are "lifted" into 3D using an estimated depth and then "splatted" onto the same BEV grid, where the two feature maps are concatenated and fused.

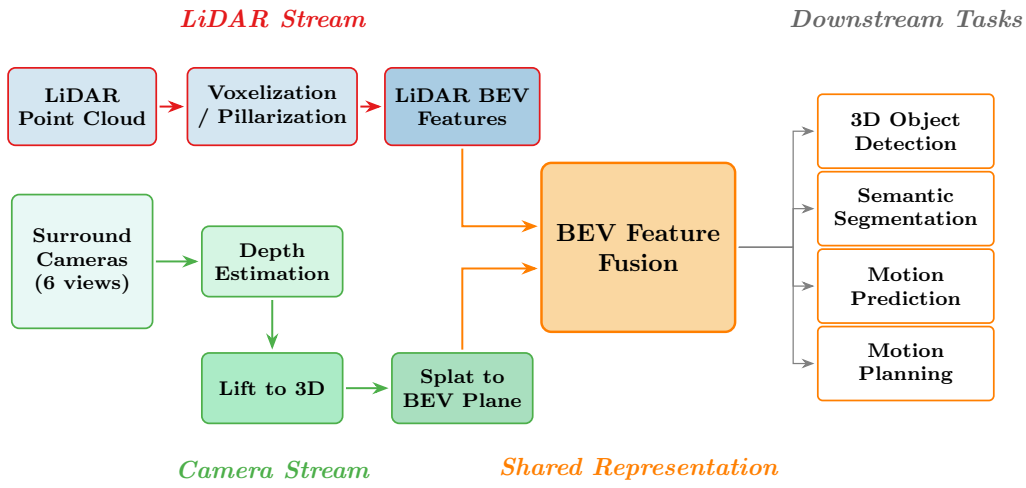

Figure 4. The canonical BEV fusion pipeline. LiDAR stream (top): point cloud → voxelization/pillarization → LiDAR BEV features. Camera stream (bottom): multi-view images → depth estimation → lift to 3D → splat to BEV plane → camera BEV features. The two streams are fused in the shared BEV grid, which serves as the input to multiple downstream task heads (detection, segmentation, prediction, planning).

BEV methods can be grouped by their primary technical contribution:

Lifting mechanisms. Lift–Attend–Splat [76] deliberately removes explicit depth estimation from the camera-to-BEV transform: instead of predicting a per-pixel depth distribution (as in Lift–Splat–Shoot, LSS), it uses a transformer to attend directly between camera columns and BEV rays, letting cross-attention with the LiDAR BEV features implicitly resolve where each camera feature belongs. The three stages are (1) *lift* image features to a bird's eye ray, (2) *attend* across modalities for selective fusion, and (3) *splat* onto the BEV grid. Removing the depth-prediction bottleneck makes the method robust to the depth-supervision noise that degrades LSS-style lifting when LiDAR is sparse.

Spatio-temporal learning. BEVFormer [30] introduced spatio-temporal transformers that learn BEV representations by attending to relevant image regions across both space and time. Although BEVFormer itself is camera-only, its deformable spatio-temporal attention became a template for subsequent LiDAR–camera BEV methods, which add the point cloud as a geometric anchor. BEVFusion4D [49] extended this concept with 4D template up-sampling, making it particularly effective for moving objects where the geometric prior from LiDAR resolves ambiguities in camera-based motion estimation.

Task unification and planning integration. BEV-CAM3D [77] unified BEV representation learning across multiple perception tasks (detection, segmentation, tracking) within a single architecture. BevDrive [78] pushed BEV fusion further downstream by integrating perception with motion planning in a shared BEV space, demonstrating the natural synergy between representation and decision-making.

Architectural efficiency. Simple-BEV [79], in the camera(+radar) setting, showed that a plain BEV architecture without complex attention matches far more elaborate designs once input resolution and batch size are properly tuned, evidence that implementation details often outweigh architectural novelty. The LiDAR–camera architecture of [80] (confusingly also named SimpleBEV, but an unrelated later work) and CL-fusionBEV [81] carry the same lesson into explicit LiDAR–camera fusion, demonstrating that carefully designed fusion modules and hybrid attention mechanisms matter more than raw model complexity.

Table 7 provides a structured comparison of the main BEV methods discussed above, highlighting the diversity of design choices within this paradigm. All performance values are as reported in the respective original publications on the nuScenes validation set (unless otherwise noted); rows are grouped by evaluation split, and comparisons are drawn only within a group.

Table 7 mixes validation and test results, which are not numerically comparable: the gap between the two splits is not systematic, since test submissions may use a different training protocol and can score higher than validation (Section 4.3). The observations below are therefore drawn only within the same split. First, the camera-only BEVFormer baseline (48.1 mAP/56.9 NDS on test) sits well below every LiDAR–camera entry, quantifying the gain that the LiDAR modality contributes within an otherwise comparable BEV transformer design. Second, within the validation group, the methods employing cross-attention for modality interaction (BEV-CAM3D, CL-fusionBEV) report NDS 2–3 points above Simple-BEV, which relies on rectified depth without cross-modal attention. This is consistent with the view that how LiDAR and camera features interact matters, but the margin is small and the methods differ in other design choices, so it should be read as an indication rather than an established result. No conclusion on the value of temporal modeling can be drawn from the table, since the only temporal fusion method (BEVFusion4D) is evaluated on a different split. Third, Lift–Attend–Splat and BevDrive represent divergent philosophies (the former challenges the necessity of explicit depth estimation, while the latter extends BEV beyond perception into planning), but neither reports standard detection benchmarks, making direct quantitative comparison impossible.

Table 7. Comparison of BEV fusion methods by technical approach and reported benchmark performance on nuScenes. All values are from the respective original publications. "—" indicates that the metric was not reported in the original paper. Rows are grouped by evaluation split (validation set, test server, not reported); values from different groups are not directly comparable.

Method	Temporal	mAP	NDS	Approach/Key Idea
Validation set				
BEV-CAM3D [77]	No	73.9	76.2	Deformable cross-modal attention; unified multi-task BEV head
CL-fusionBEV [81]	No	73.3	75.5	Implicit view transform with cross-attention for LiDAR–camera BEV interaction
SimpleBEV [80]	No	71.4	73.5	Cascade + LiDAR-rectified depth; improved encoder and auxiliary camera branch
Test server				
BEVFusion4D [49]	Yes (4D)	73.3	74.7	LiDAR-guided view transformer; 4D templates + temporal deformable alignment
BEVFormer [30] (camera-only baseline)	Yes	48.1	56.9	Spatio-temporal deformable attention over learnable BEV queries
Detection benchmark not reported				
Lift–Attend–Splat [76]	No	—	—	Attention-based feature selection; bypasses explicit depth estimation
BevDrive [78]	Yes	—	—	BEV projection with a planning head; perception–planning unification

7.3. Surround-View + 3D LiDAR

The full autonomous driving sensor suite (6 surround cameras + 1 LiDAR 360° as in nuScenes, or 5 cameras + 5 LiDARs as in Waymo) represents the high end of the depth continuum. Here, the challenge shifts from "how to get enough depth" to "how to efficiently fuse abundant multi-view, multi-modal information."

Real-time object detection using LiDAR and camera fusion [35] demonstrated a practical end-to-end approach: projecting 3D LiDAR points into 2D camera views for joint processing, achieving real-time inference suitable for deployment. Multi-sensor fusion and segmentation approaches [36] handle the full pipeline from raw sensor data to object detection and segmentation. Recent work on multi-modal fusion object detection algorithms [82,83] has further advanced the state of the art. The addition of 4D radar in RAF [72] (Section 6) points in the same direction, toward reliability-weighted multi-sensor architectures.

7.4. High-End (128+ Beams)

At 128+ beams, the LiDAR point cloud approaches the density of a depth map, raising a provocative question: is camera fusion still necessary? The answer emerging from the surveyed literature is yes, but the role of the camera shifts from depth provider to semantic specialist. While 128-beam LiDAR can detect object shapes with high fidelity, it cannot distinguish a stationary car from a wall, a traffic light color, or a lane marking. The camera provides this essential semantic context.

The fusion strategy for ultra-dense LiDAR is evolving toward geometry-guided semantic extraction: the dense LiDAR geometry defines where to look, and the camera defines what is being looked at. Attention mechanisms that use depth as a query bias are the natural candidates in this regime [51,54], although their published results come from 32-beam nuScenes data (Section 4.3).

8. Depth Completion: Bridging Sparse Measurements with Dense Appearance

Depth completion (predicting a dense depth map from sparse LiDAR measurements guided by a corresponding RGB image) is arguably the most versatile mechanism in the LiDAR–camera fusion toolbox. It serves as the critical bridge across the depth continuum, providing sparse-LiDAR systems with the dense depth they lack, provided that it is trained at the target density (as discussed below).

8.1. Problem Formulation

Given a sparse depth map $\mathbf{S} \in \mathbb{R}^{H \times W}$ (where most pixels have no valid depth) and an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, depth completion aims to predict a dense depth map $\hat{\mathbf{D}} \in \mathbb{R}^{H \times W}$ where every pixel has a valid depth value. The key challenge is that valid depth pixels from LiDAR are both sparse (typically <5% density for 16-beam, ∼10–15% for 64-beam, when projected onto a high-resolution image) and non-uniformly distributed (dense near the sensor, sparse at range). Figure 5 illustrates the depth completion pipeline.

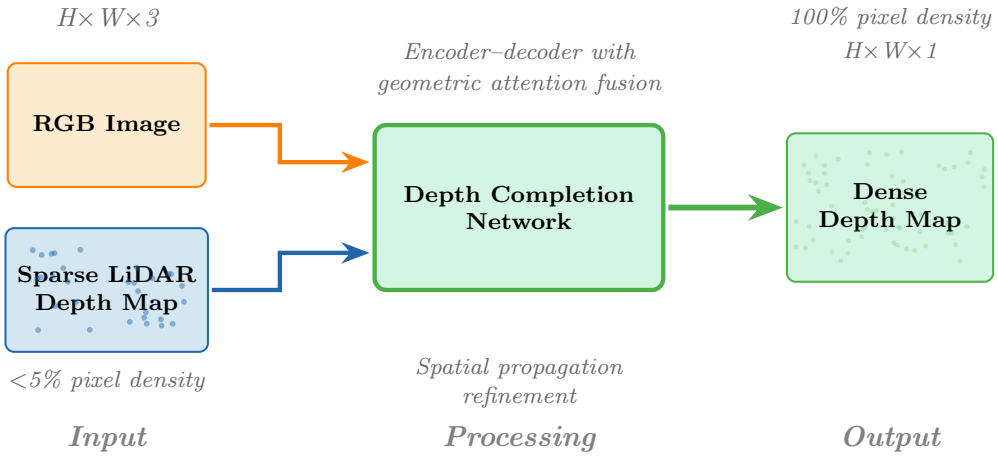

Figure 5. The depth completion pipeline. **Input:** a sparse LiDAR depth map (<5% density) paired with a dense RGB image. The depth completion network (encoder–decoder with geometric attention) propagates sparse depth values guided by appearance cues from the RGB image, producing a dense depth map (100% density). Key methods include MSMDFusion, DMFusion, FastFusion, GAC-Net, and foundation model distillation approaches.

8.2. From Classical to Foundation Models

Early depth completion methods used classical image processing: bilateral filtering to propagate sparse depth values to neighboring pixels based on color similarity, and convex optimization with smoothness priors. The introduction of convolutional neural networks substantially improved performance. Sparse-to-Dense [84] was among the first to frame depth completion as an end-to-end learning problem, predicting a dense map directly from sparse depth samples and a single RGB image with a standard encoder–decoder network. Convolutional spatial propagation network (CSPN) [85] improved on this by learning a pixel-wise affinity matrix that iteratively propagates depth values through a recurrent, spatially local diffusion process, better preserving object boundaries than direct regression.

Modern approaches leverage attention mechanisms and foundation models. GAC-Net [45] introduces geometric-attention fusion that explicitly models 3D geometric priors when combining RGB and sparse depth, achieving state-of-the-art performance on the KITTI depth completion benchmark. The distillation of monocular foundation models for depth completion [25] represents a change of approach: rather than training from scratch, a two-stage knowledge distillation framework transfers the dense depth knowledge from large pre-trained monocular depth models to guide sparse-to-dense completion.

8.3. Stereo-LiDAR Depth Fusion

Stereo cameras occupy a unique position in the depth continuum: they produce dense but accuracy-limited depth, which can be refined with sparse but precise LiDAR anchors. FastFusion [27] proposed a three-stage deep fusion scheme integrating LiDAR priors into every step of classical stereo matching: cost computation, cost aggregation, and disparity refinement, achieving real-time high-precision dense depth.

Fan et al. [86] introduced a holistic evidential fusion framework that quantifies the uncertainty of depth estimates from both stereo and LiDAR, enabling the fusion to weight each modality's contribution based on its reliability at each pixel. Similarly, probabilistic multi-modal depth estimation [33] addresses the inherent uncertainty in sensor measurements through Bayesian fusion frameworks. This uncertainty-aware approach is particularly valuable for safety-critical applications where knowing when depth estimates are unreliable is as important as the estimates themselves.

8.4. Scaling with LiDAR Density

A critical open question is how depth completion performance scales with LiDAR density. Park et al. [73] explicitly studied this, evaluating depth completion methods across a range of input densities from 4-line LiDAR (ultra-sparse) to 64-line (standard). They demonstrated that most methods trained on 64-line data degrade severely when tested on sparser inputs.

For low-cost platforms, this means that training and evaluating depth completion methods at the specific LiDAR density of the target deployment is essential, and methods that explicitly model sparsity (rather than treating it as an invariant) show better cross-density generalization [73].

8.5. Real-Time Depth Completion

For autonomous systems operating under real-time constraints, the computational cost of depth completion is a critical factor. FastFusion [27], described above, achieves real-time performance (>30 fps). The real-time depth completion network of Chen et al. [46] avoids costly 3D convolutions entirely, using injection guidance and kernel-connected spatial propagation for efficient refinement. Unsupervised approaches [71] offer another route to efficiency by removing the need for dense ground-truth depth during training.

9. Calibration and Synchronization: The Hidden Foundation

Sensor fusion is only as good as the calibration that enables it. Even the most sophisticated fusion architecture will produce incorrect results if LiDAR points are projected onto the wrong image pixels due to calibration errors, or if temporally misaligned data from different sensors are fused.

9.1. From Calibration to Fusion: The Perception Pipeline

In a deployed system, calibration and fusion form a single chain: offline (target-based) calibration provides the initial extrinsics; online refinement compensates for drift; temporal synchronization and ego-motion compensation (de-skewing of the LiDAR sweep, which

is acquired over 50–100 ms while the platform moves) bring both streams to a common timestamp; the calibrated transformation maps the data into a shared frame; cross-sensor correspondences are established; and only then are features fused and passed to the perception head. Every fusion strategy relies on the same geometric core. A LiDAR point $\mathbf{p}_L$ is mapped to the camera frame and then to the image plane as

$$\mathbf{p}_C = \mathbf{R}\,\mathbf{p}_L + \mathbf{t}, \qquad z_C \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}\,\mathbf{p}_C, \tag{1}$$

where $(\mathbf{R}, \mathbf{t})$ are the extrinsics, $\mathbf{K}$ is the camera intrinsic matrix with focal length f (in pixels), and z_C is the depth of the point in the camera frame. Equation (1) also shows how errors propagate. A small rotation error $\Delta\theta$ shifts the projected point by approximately $f\,\Delta\theta$ pixels, independently of distance (about 20 px for $\Delta\theta = 1^\circ$ and $f \approx 1200$ px, typical of automotive cameras); a translation error $\|\Delta\mathbf{t}\|$ shifts it by approximately $f\,\|\Delta\mathbf{t}\|/z_C$, which matters most for nearby objects; and a residual time offset τ between the two streams displaces an object moving at speed s by $s\,\tau$ (1 m for an object moving at 20 m/s with $\tau = 50$ ms). The fusion strategies of Section 4 differ in where Equation (1) is applied and how hard the resulting correspondence is, which in turn determines how calibration errors reach the perception output (Table 8).

Table 8. How each fusion strategy uses the calibration: the frame in which the modalities are fused, how cross-sensor correspondence and feature association are established, and the effect of extrinsic or temporal misalignment on the downstream output.

Strategy	Fusion Frame	Correspondence and Feature Association	Use of Calibration	Effect of Misalignment
Projection-based	Image plane	Hard: each point takes the semantics or features of the pixel it projects onto	Equation (1) applied to every point	Points inherit wrong semantics, especially at object boundaries
Depth completion	Image plane	Hard: projected sparse depth map used as input to an image-guided network	Equation (1) to build the sparse depth map	Depth edges shifted from image edges; errors spread to the densified map
BEV fusion	BEV grid (ego frame)	Implicit: LiDAR voxelized directly; camera features lifted with predicted depth and placed in the same grid	Camera-to-ego transform to place lifted features	Camera and LiDAR features land in different cells; among the most affected tightly coupled designs (Figure 6)
Attention-based	Feature/query space	Soft: projected reference points guide where to sample, the attention weights learn the final association	Initial guide only	Graceful degradation; learned association absorbs small errors
Late fusion	Ego/world frame	Box-level: per-sensor detections matched by overlap or distance (e.g., Hungarian assignment)	Only to express boxes in a common frame	Missed or wrong box associations; per-sensor detections unaffected

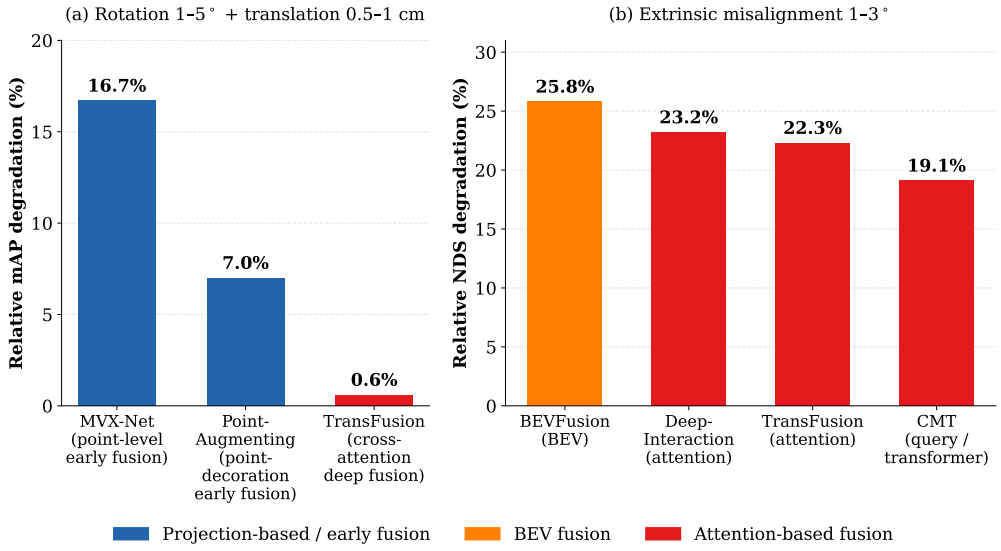

Figure 6. Measured impact of camera–LiDAR extrinsic misalignment on detection performance, from the two robustness benchmarks discussed above. (**a**) Relative mAP degradation under combined rotation (1–5°) and translation (0.5–1 cm) offsets on nuScenes. (**b**) Relative NDS degradation under 1–3° extrinsic misalignment, computed as 1 − RA from the reported Resistance Ability. Bars are colored by fusion family. Absolute values are not comparable across the two panels (different metrics and corruption protocols); the consistent finding is that point-level/early fusion is the most calibration-sensitive and attention/query-based fusion the least.

9.2. Spatial Calibration (Extrinsic)

Extrinsic calibration determines the rigid-body transformation (rotation **R** and translation **t**) between the LiDAR and camera coordinate frames. Traditional methods use calibration targets (typically checkerboard patterns visible to both sensors), requiring manual data collection in controlled settings. While accurate, these methods are impractical for large-scale deployment, where sensors may drift over time due to vibration and temperature changes.

Learning-based and targetless methods have gained significant attention as alternatives. Several comprehensive reviews have surveyed the calibration landscape: the review of extrinsic parameter calibration for LiDAR and camera [87], the survey on external multi-modal imaging sensor calibration [88], the survey of deep learning-based extrinsic calibration [89], and the survey of extrinsic calibration on LiDAR–camera systems for intelligent vehicles [90]. BEV-based calibration [91] represents a particularly elegant approach: by aligning the BEV features produced independently from LiDAR and camera data, the calibration matrix can be estimated as the transformation that maximizes feature correspondence, a process that can be integrated directly into the perception pipeline. CalibRefine [40] follows a similar online, targetless philosophy: it detects common features in both modalities, estimates a coarse homography, and refines it iteratively as frames accumulate, with a final Vision Transformer cross-attention stage that corrects residual non-planar misalignment. Notably, it frames calibration explicitly as a service to fusion rather than to 3D reconstruction. Joint intrinsic and extrinsic calibration [92] tackles the interdependent nature of these parameters.

9.3. Temporal Synchronization

Temporal synchronization ensures that data from different sensors correspond to the same moment in time. LiDAR typically operates at 10–20 Hz, while cameras run at 30–60 Hz, creating a fundamental mismatch. Hardware synchronization using Global Positioning

System (GPS) pulse-per-second (PPS) signals provides the most accurate solution [93], but adds cost and complexity.

Software-based approaches use timestamp interpolation or, increasingly, learned robustness to temporal misalignment. The TimeAlign method [94] explicitly models the temporal offset between sensor streams as part of the detection pipeline, training the network to be robust to realistic timing errors rather than requiring perfect synchronization. Multi-sensor calibration frameworks integrating LiDAR, camera, and an inertial measurement unit (IMU) [95] address the practical challenges of full-suite synchronization. [96] instead proposes a hardware-trigger approach based on the Precision Time Protocol (PTP), synchronizing LiDAR and multiple cameras to a common master clock with adaptively estimated trigger delays, without relying on GPS.

9.4. Impact of Calibration Quality

The sensitivity of fusion performance to calibration errors varies significantly across fusion strategies, forming an approximate robustness ordering:

1. **Projection-based (early fusion):** Most sensitive. A 1-degree rotation error can displace projected LiDAR points by about twenty pixels on the image plane (approximately $f\,\Delta\theta$, i.e., proportional to the focal length but nearly independent of object distance; translation errors instead decrease with distance, Section 9.1), causing feature-level fusion to combine semantically unrelated information. This strategy requires the highest calibration quality and is least suitable for applications where sensor drift is expected.
2. **Middle and BEV fusion:** Intermediate, but architecture-dependent. Volumetric representations provide some tolerance to individual point displacement (voxels and pillars average point features within spatial bins), yet BEV methods that lift camera features with a learned depth remain sensitive to errors in the camera–LiDAR transform: in the MultiCorrupt benchmark, BEVFusion is in fact the least robust of the tightly coupled architectures tested. Cross-attention mechanisms, which learn soft rather than hard correspondences, are consistently more robust than either.
3. **Late fusion:** Expected most robust. Since each sensor's detection pipeline operates independently on its native coordinate frame, calibration errors affect only the final bounding box association step, a significantly lower-dimensional problem than pixel-level or feature-level alignment. No large-scale robustness benchmark currently includes a late-fusion detector, so this expectation rests on architectural reasoning rather than measurement.

This qualitative ordering is broadly consistent with the two published robustness benchmarks that explicitly perturb the camera–LiDAR extrinsics, summarized in Figure 6. Yu et al. [97] apply combined rotation (1–5°) and translation (0.5–1 cm) offsets on nuScenes: hard point-level early fusion degrades sharply (MVX-Net loses 16.7% relative mAP), point-decoration early fusion is intermediate (PointAugmenting, −7.0%), and cross-attention deep fusion is almost unaffected (TransFusion, −0.6%). The MultiCorrupt benchmark [74], applying 1–3° extrinsic misalignment, shows that among tightly coupled architectures the BEV-level fusion of BEVFusion degrades most (−25.8% relative NDS), while query/transformer fusion (CMT) degrades least (−19.1%). The two benchmarks disagree on the absolute magnitude for a given method, because their corruption protocols differ, but agree on the direction: robustness to extrinsic error increases as the fusion mechanism moves from hard geometric projection toward learned soft association.

This calibration sensitivity ordering should be explicitly factored into fusion strategy selection: applications where frequent recalibration is impractical (e.g., agricultural robots operating in rough terrain, long-duration drone missions) should favor late or

attention-based fusion over hard-projection early fusion, and budget for online extrinsic self-calibration where early fusion is unavoidable.

10. Robustness to Adverse Conditions and Sensor Failures

10.1. Weather and Lighting Degradation

Adverse weather conditions (rain, fog, snow, and low light) degrade both LiDAR and camera performance, but through different mechanisms. LiDAR suffers from backscattering (laser pulses reflecting off rain droplets or fog particles before reaching the target) and signal attenuation (reduced range), producing noisy and incomplete point clouds [98]. Cameras degrade through reduced contrast, glare from wet surfaces, and complete loss of features in darkness.

The complementary failure modes of the two sensors create an opportunity: fusion can provide robustness if the system knows when to trust which modality. The uncertainty-aware multi-modal fusion (UAMF) framework [37] integrates LiDAR, RGB, and weather priors through dynamic cross-modal attention and Bayesian uncertainty estimation, adjusting modality weights based on environmental conditions.

SAMFusion [38] extends the sensor suite to include gated cameras (near-infrared (NIR)) and radar alongside standard LiDAR and RGB cameras, demonstrating that sensor diversity is the most effective defense against environmental degradation. FogFusion [39] specifically targets foggy conditions with a synergistic camera–LiDAR fusion that models and compensates for fog-induced attenuation. The AWAD-Fusion framework [99] introduces dynamic multi-sensor fusion that adapts its fusion strategy based on real-time weather assessment. The quantitative impact of adverse conditions on perception performance follows consistent patterns across multiple studies, as summarized in Table 9. Rather than reporting single-point accuracy figures (which vary significantly across datasets, sensor models, and evaluation protocols), the table captures the relative degradation and complementary failure modes that consistently emerge from the literature.

Table 9. Qualitative impact of adverse conditions on LiDAR–camera fusion and single-modality perception. Symbols: ■ = near-baseline performance (<10% degradation); ▲ = moderate degradation (10–30%); ♦ = severe degradation (30–50%); × = critical failure (>50% degradation). Patterns are compiled from multiple independent studies with per-row references.

Condition	LiDAR-Only	Camera-Only	Fused	Primary Degradation Mechanism	Refs.
Clear/Day (baseline)	■	■	■	—	standard
Light rain	▲	▲	■	LiDAR: droplet backscattering; Camera: reduced contrast	[37,100]
Heavy rain	♦	♦	▲	LiDAR: severe attenuation + splash; Camera: lens occlusion	[100,101]
Fog (vis. <100 m)	×	♦	▲	LiDAR: Mie scattering of laser pulses	[39,100]
Snow/snowfall	♦	♦	▲	LiDAR: snowflake false returns; Camera: whiteout	[100,102]
Night/Low light	■	×	▲	Camera: signal-to-noise ratio collapse; LiDAR resilient (active sensor)	[38,102]
Direct sun glare	■	×	▲	Camera: pixel saturation; LiDAR unaffected	[102,103]

Two key observations emerge from Table 9. First, the complementarity of failure modes is systematic rather than incidental: LiDAR fails primarily through atmospheric scattering (fog, heavy precipitation), while cameras fail primarily through lighting extremes (night, glare, whiteout). No single condition simultaneously disables both modalities at the critical-failure level, supporting the fundamental premise of multi-modal fusion for robustness. Second, fusion provides the largest relative gain precisely where single-modality degradation is most asymmetric. At night, where LiDAR maintains near-baseline performance while cameras collapse, fusion nearly recovers baseline performance by

relying predominantly on the LiDAR stream. This observation has practical implications for fusion architecture design: in conditions with strongly asymmetric degradation, dynamic modality weighting (as in UAMF [37]) has been shown to outperform static fusion schemes that combine modalities with fixed weights regardless of environmental state.

10.2. Sensor Failure Modes and Graceful Degradation

A critical but understudied aspect of fusion is behavior under partial sensor failure. Sensor failures in autonomous systems are not rare: cameras can be temporarily blinded by direct sunlight or occluded by dirt and insects on the lens; LiDAR can be saturated by retro-reflective surfaces (road signs, license plates) or produce erroneous returns from highly specular objects (wet road surfaces, glass facades). When one sensor fails, the fusion system should degrade gracefully rather than catastrophically.

The architecture of the fusion pipeline largely determines the failure response. In early and middle fusion, where feature representations from both modalities are intermixed before detection, corrupted features from a failing sensor can propagate through the shared representation, potentially degrading the output below what either sensor would produce alone. BEV architectures that rely on LiDAR for grid construction and cameras for semantic content are particularly vulnerable: if the LiDAR fails, the BEV grid itself becomes unreliable; if cameras fail, the semantic content of grid cells is compromised. Mitigation strategies include confidence-weighted feature fusion (assigning per-sensor reliability scores to each feature), sensor health monitoring with online anomaly detection, and fallback pathways that route single-modality detections when cross-modal consistency is low.

Late fusion architectures follow a fundamentally different failure mode: because detection pipelines operate independently, a sensor failure reduces to a missing-modality scenario rather than a corrupted-feature scenario. The system can fall back to single-modality detections from the healthy sensor without any architectural modification, a property known as *graceful modality degradation*. This robustness, however, is hypothesized rather than measured: it follows from the architecture, but no large-scale robustness benchmark includes a late-fusion detector (Section 9). The closest evidence is indirect: in MultiCorrupt [74], SparseFusion, whose modality-specific branches work largely independently, ranks second of five detectors, but first is CMT, a tightly coupled detector trained with randomly missing modalities, so robustness can also come from training rather than from decoupling. Moreover, it holds only for sensor loss: a degraded but still active sensor (e.g., producing spurious returns) passes its false detections to the output unless the association stage checks cross-modal consistency. This robustness to sensor loss is the primary reason late fusion remains relevant despite its lower nominal accuracy. A recent investigation [34] specifically studied this vulnerability in monocular–LiDAR fusion and confirmed that introducing a stereo baseline provides built-in depth redundancy, so that if one camera fails, the surviving monocular stream combined with LiDAR can maintain partial functionality.

The trade-off is therefore plausible but not yet quantified: architectures optimized for peak performance under nominal conditions (early/middle/BEV fusion) appear to sacrifice failure robustness that modular late-fusion designs provide by construction, unless they are explicitly trained for missing modalities. Practical deployment decisions should weigh this trade-off based on the safety-criticality of the application and the expected sensor failure rate in the target operating environment.

10.3. Cross-Domain Generalization

The gap between training datasets (recorded in a few cities with mostly favorable weather: KITTI in Karlsruhe, nuScenes in Boston and Singapore, Waymo mainly in Phoenix,

Mountain View, and San Francisco) and deployment environments (different cities, weather, traffic norms) remains a significant challenge. Domain adaptation for multi-modal perception is an emerging area, with cross-dataset late fusion [59] suggesting that modular late-fusion frameworks can be more transferable across domains than tightly coupled middle-fusion architectures.

11. Datasets, Benchmarks, and Evaluation

11.1. Standard Benchmarks

The three dominant datasets for LiDAR–camera fusion research are summarized in Table 10. Scenes denotes the number of distinct, annotated driving sequences (short continuous recordings, typically a few seconds to tens of seconds long) that make up the dataset; 3D Boxes is the total number of 3D bounding-box annotations across all scenes, one per labeled object instance per frame; Classes is the number of distinct semantic object categories annotated (e.g., car, pedestrian, cyclist). Note that KITTI reports its scale in individual annotated frames rather than multi-frame scenes, so its "7.5K frames" is not directly comparable to the scene counts of nuScenes and Waymo.

Table 10. Comparison of major autonomous driving datasets for LiDAR–camera fusion.

Dataset	Sensors	Scenes	3D Boxes	Classes	Key Feature
KITTI [16]	1 × 64-beam LiDAR, 2 × stereo	7.5K frames	80 K	8	Pioneer; stereo available
nuScenes [17]	1 × 32-beam LiDAR, 6 × cam, 5 × radar	1K scenes (20 s)	1.4 M	23	Full sensor suite
Waymo [18]	5 × LiDAR, 5 × cam	1.15K scenes	12 M	4	Highest annotation density

KITTI [16] established the foundation for 3D object detection benchmarks. Its 64-beam LiDAR + stereo camera setup represents the mid-range of our depth continuum. nuScenes [17] introduced the full sensor suite paradigm with 6 cameras, 1 LiDAR (32-beam), and 5 radars, enabling surround-view perception research. Waymo Open Dataset [18] provides the highest annotation density with 12M+ 3D boxes across 1150 scenes. Beyond these mainstream benchmarks, RTPSeg [104] provides synchronized RGB, infrared (IR), and LiDAR frames with pointwise annotations covering day and night conditions, while the ROVR Open Dataset [105] targets dense 3D point clouds for depth estimation and semantic segmentation in robotics contexts.

11.2. A Critical Gap: Missing 2D LiDAR Benchmarks

A significant finding of this review is the near-total absence of standardized benchmarks for 2D LiDAR–camera fusion. All major datasets assume at minimum a 32-beam 3D LiDAR, creating a disconnect between the academic literature (which assumes 3D LiDAR) and the practical reality of low-cost robotics (which predominantly uses 2D LiDAR). Furthermore, no major dataset covers non-automotive domains such as agriculture, warehouse logistics, or indoor service robotics with multi-modal annotations at the scale of the driving benchmarks, despite these domains representing a large and growing share of autonomous system deployments. Section 11.3 reviews the public datasets that pair a 2D LiDAR with a camera and their limitations, and Section 12 outlines the benchmark needed to close this gap.

11.3. Data Availability Across Application Domains

While the autonomous driving domain benefits from large-scale, richly annotated datasets, other application domains covered in this review suffer from a severe data

scarcity. Table 11 summarizes representative datasets across the four application domains, highlighting the stark imbalance in data availability.

Table 11. Representative datasets across application domains for LiDAR–camera fusion. For the three non-automotive domains, the table lists public datasets that pair a 2D or low-beam LiDAR with a camera; none of them is a fusion benchmark (Section 11.2), and most papers in these domains still evaluate on custom setups, i.e., small, typically proprietary or single-group rigs [23,28,61].

Domain	Dataset	Sensors	Scale	Ground Truth
Auton. Driving	KITTI [16]	64-beam LiDAR, stereo	7.5K frames	3D boxes
	nuScenes [17]	32-beam LiDAR, 6× cam, 5× radar	1K scenes	3D boxes
	Waymo [18]	5× LiDAR, 5× cam	1.15K scenes	3D boxes (4 classes)
	RTPSeg [104]	RGB, IR, LiDAR	3K+ frames	Point-wise semantic labels
AGV/Logistics	TorWIC [106]	2D LiDAR, RGB-D	18 trajectories	Poses from 2D LiDAR SLAM; no object labels
	KIT intralogistics [107]	2 × 2D LiDAR, RGB	74 frames	Instance masks on images (13 classes)
Agriculture	Sugar Beets 2016 [108]	16-beam and 3D laser scanners, RGB-D, multispectral camera	3 months, ∼5 TB	Pixel-wise plant/weed labels on images
Indoor Robots	Robot@Home [109]	2D LiDAR, 4 × RGB-D	5 homes, 36 rooms	Per-pixel object and room labels; labeled 3D reconstructions
	OpenLORIS-Scene [110]	2D LiDAR, RGB-D, stereo fisheye	5 scenes, 22 sequences	Poses (the 2D LiDAR is used to generate them)
	JRDB [111]	2 × 2D LiDAR, 2 × 16-beam, 360° stereo, RGB-D	64 min, 54 sequences	2D/3D boxes of people, on images and 16-beam clouds only
	MIT Stata Center [112]	2D LiDAR, stereo, RGB-D	38 h, 42 km	Poses only

Some public datasets do pair a 2D LiDAR with a camera (Table 11), but none of them provides a benchmark for 2D LiDAR–camera fusion. Most were recorded for SLAM or long-term mapping and provide only ground-truth poses: the MIT Stata Center dataset [112], OpenLORIS-Scene [110], where the 2D LiDAR is itself used to generate the ground truth, and the TorWIC warehouse dataset [106]. Robot@Home [109] adds per-pixel object and room labels, but covers only five apartments. JRDB [111] carries two 2D LiDARs on a social robot, yet its person annotations refer to the 16-beam point clouds and the images; labels for the 2D scans have to be derived from them, as done by Jia et al. [113], who train 2D LiDAR person detectors with pseudo-labels generated by a calibrated camera. Recent releases are either very small, such as a 74-frame intralogistics set with RGB images, 2D scans, and instance masks [107], or based on 2D LiDAR alone, such as Semantic2D [114]. What is missing is therefore not data as such, but data with a common task, annotations suited to fusion, and an evaluation protocol.

The automotive datasets carry their own restrictions: they are geographically narrow (Section 10) and, in Waymo's case, annotate only four object classes. In the other three domains, most papers evaluate on custom setups, which share three limitations. First, there is no common evaluation protocol: each work defines its own data split, metrics, and matching thresholds, and often its own baselines. Second, 3D ground truth is rarely available: annotations, when present, are 2D boxes or masks on the images, or robot poses,

since labeling 3D boxes without a dense LiDAR is expensive. Third, the recordings cover little variation: usually one building or field over a short period, under indoor or daytime lighting, and rarely at night, in rain, or in dust. Moreover, data and calibration files are seldom released, so the results cannot be reproduced.

The absence of non-automotive benchmarks has two concrete consequences. First, it prevents fair comparison between low-cost fusion methods, as each research group evaluates on its own custom setup. Second, it discourages the computer vision community from developing fusion methods tailored to the specific challenges of agricultural, logistics, and indoor domains, challenges that differ substantially from autonomous driving (e.g., deformable objects in agriculture, reflective surfaces in warehouses, dynamic human environments indoors); a multi-density, multi-domain benchmark that would address both issues is outlined in Section 12.

11.4. Evaluation Metrics

The quantitative comparison of LiDAR–camera fusion methods relies on a set of standardized metrics that capture both detection accuracy and spatial alignment quality. Understanding these metrics and their limitations is essential for interpreting the benchmark results reported throughout this review.

3D Object Detection Metrics. The primary metric is mean Average Precision (mAP) computed over 3D Intersection-over-Union (IoU) thresholds. A 3D bounding box prediction is considered a true positive if its IoU with the ground truth exceeds a threshold, typically 0.7 for cars (KITTI) or 0.5–0.7 with class-specific thresholds (nuScenes). The KITTI benchmark evaluates at three difficulty levels based on object size, occlusion, and truncation: *easy* (fully visible, minimum bounding box height of 40 px), *moderate* (partially occluded, minimum height of 25 px), and *hard* (heavily occluded, minimum height of 25 px). The nuScenes benchmark introduced the nuScenes Detection Score (NDS), a composite metric defined as

$$\text{NDS} = \frac{1}{2}\left[\text{mAP} + \frac{1}{5}\sum_{e\in\mathcal{E}}(1 - \min(1, \text{mTP}_e))\right] \tag{2}$$

where $\mathcal{E} = \{\text{translation, scale, orientation, velocity, attribute}\}$ is the set of true positive error metrics, and mTP_e is the mean true positive metric for error type e. NDS explicitly penalizes methods that achieve high mAP with poor localization, making it a more holistic measure of detection quality than mAP alone.

Depth Completion Metrics. The standard metrics, computed over the N pixels with ground-truth depth d_i^{gt} and predicted depth $\hat{d}_i$, are as follows:

- **RMSE** (Root Mean Square Error): $\text{RMSE} = \sqrt{\frac{1}{N}\sum_i (\hat{d}_i - d_i^{\text{gt}})^2}$, capturing large errors.
- **MAE** (Mean Absolute Error): $\text{MAE} = \frac{1}{N}\sum_i |\hat{d}_i - d_i^{\text{gt}}|$, robust to outliers.
- **iRMSE** (Inverse RMSE): $\text{iRMSE} = \sqrt{\frac{1}{N}\sum_i (1/\hat{d}_i - 1/d_i^{\text{gt}})^2}$, which weights near-range errors more heavily.
- δ_t **threshold accuracy**: Percentage of pixels where $\max(\hat{d}_i/d_i^{\text{gt}}, d_i^{\text{gt}}/\hat{d}_i) < t$, typically reported at $t \in \{1.25, 1.25^2, 1.25^3\}$.

Metrics for Low-Cost Robotics. mAP and NDS assume 3D bounding boxes, and NDS also velocities and attributes defined for traffic participants; RMSE and MAE assume a dense or semi-dense depth reference. Most 2D LiDAR setups have neither, and these metrics do not measure what such robots are built for. Task-specific metrics are more appropriate:

- **2D SLAM**: Trajectory and map accuracy, usually reported as absolute trajectory error (ATE) and relative pose error (RPE) [115]. They only require ground-truth poses, the kind of ground truth that most of the robotics datasets in Table 11 actually provide.

- **Detection and tracking with a 2D LiDAR**: A distance threshold on the ground plane replaces the 3D IoU of mAP, as in person detection from 2D range data, where a detection counts as correct within 0.3 or 0.5 m of a ground-truth person [113]. The same distance-based matching can be used to compute standard multi-object tracking metrics, such as the number of identity switches between consecutive frames.
- **Semantic mapping**: Per-class IoU between predicted and annotated cells or points, on the map or on the scan, as in 2D LiDAR semantic segmentation [114].
- **Efficiency**: Since these systems run on embedded hardware, accuracy should be reported together with latency and power consumption.

A fusion method for low-cost robots should therefore be judged by whether it improves these task metrics over LiDAR-only and camera-only baselines, rather than by mAP or NDS.

Limitations of Current Metrics. Standard metrics have known blind spots: mAP with fixed IoU thresholds does not capture detection quality for small or distant objects well; NDS aggregates diverse error types into a single scalar, potentially masking trade-offs; and depth completion metrics are typically computed over the full image, including regions far from any LiDAR measurement where errors are largest but least consequential for downstream tasks. The development of metrics that better capture real-world deployment requirements (such as latency-aware metrics, calibration-robustness scores, and domain-shift resilience indices) remains an open challenge.

12. Open Challenges and Future Directions

12.1. Toward Unified Frameworks Across the Depth Continuum

A key open challenge is the development of fusion architectures that adapt their strategy based on available depth information. Current methods are designed for a specific LiDAR density (typically 32 or 64 beams) and do not gracefully scale to sparser or denser configurations. A unified framework that automatically adjusts (using depth completion when LiDAR is sparse and attention-based fusion when it is dense) would enable a single architecture to serve applications across the full depth continuum. Early work on flexible depth completion [73] points in this direction.

12.2. Real-Time Depth Completion on Edge Hardware

Deploying depth completion on embedded platforms (NVIDIA Jetson, Qualcomm Snapdragon, or field-programmable gate arrays) with real-time constraints (>30 fps) remains an open problem, particularly for low-beam LiDAR where the completion task is more challenging. Hardware-accelerated 3D LiDAR-based object detection with BEV [24] and onboard camera–LiDAR deployment optimization [116] represent early steps toward practical edge deployment. Model compression, neural architecture search for depth completion, and hardware-aware design are active research areas.

12.3. Foundation Models for Multi-Modal Perception

Large pre-trained models have been highly successful in monocular vision tasks: DepthAnything for zero-shot monocular depth estimation, SAM (Segment Anything Model) for promptable image segmentation, and DINOv2 for self-supervised general-purpose visual features. Their success suggests a similar opportunity for multi-modal perception. A comprehensive survey of foundation models for autonomous driving [117] documents the rapid progress in this direction. Distilling foundation model knowledge for depth completion [25] is an early example. Future work may see unified foundation models that jointly reason about depth, semantics, and motion from multi-modal inputs, reducing the need for task-specific fusion architectures. The integration of vision–language–action (VLA)

models with multi-modal fusion [118] could further bridge perception, reasoning, and control in a unified framework.

12.4. Event Cameras and Novel Sensors

Event cameras (dynamic vision sensors), gated cameras (NIR with range-gated imaging), and 4D imaging radar are emerging sensor modalities that offer complementary characteristics to traditional LiDAR and cameras. Advances in radar–camera fusion for object detection [119] demonstrate the potential of integrating radar as a third modality, while occupancy perception (representing 3D space as occupied or free) is emerging as a unified representation for multi-modal fusion [120]. Event cameras provide microsecond temporal resolution and high dynamic range, making them particularly valuable for high-speed scenarios and challenging lighting. Integrating these novel sensors into the fusion taxonomy, and understanding where they sit on the depth continuum, is an open research direction.

12.5. Certification and Safety Standards

For safety-critical autonomous systems, sensor fusion must be certified against standards such as ISO 21448 [121] (safety of the intended functionality (SOTIF)) and ISO 26262 [122] (Functional Safety). Explainable fusion [123] and uncertainty quantification [37,86] are essential prerequisites for certification. The development of formal verification methods for learned fusion components remains a significant challenge.

12.6. Neural Radiance Fields and 3D Gaussian Splatting for Fusion

Neural radiance fields and 3D Gaussian splatting (3DGS) have emerged as powerful paradigms for novel view synthesis and scene reconstruction, with growing implications for LiDAR–camera fusion. These methods offer a fundamentally different approach to multi-modal representation: rather than projecting one modality onto the other, they learn a continuous, differentiable volumetric scene representation that can be rendered from any viewpoint and supervised with multi-modal data (RGB images for photometric consistency, LiDAR points for geometric accuracy). Early work in this direction includes LiDAR-NeRF-style approaches, which extend the NeRF volumetric rendering formulation to synthesize novel LiDAR range views (rather than only RGB views) so that sparse real scans can be augmented with rendered ones; street-view Gaussian splatting methods, which fit explicit, per-point 3D Gaussians to reconstruct large-scale dynamic driving scenes at real-time rendering speed; and methods that use 3DGS directly as a unified fusion backbone, where LiDAR anchors provide geometric regularization for the Gaussians and camera images provide appearance supervision. This emerging paradigm aligns naturally with the depth continuum concept: denser LiDAR configurations provide stronger geometric constraints for the neural field, while sparser configurations rely more heavily on photometric cues from cameras. Integrating NeRF/3DGS-based representations into the depth-centric taxonomy, and understanding their scaling behavior across the depth continuum, represents a promising direction.

12.7. Sim-to-Real Transfer and Long-Tail Robustness

A critical challenge for practical deployment is the gap between training distributions and real-world operating conditions. Most fusion methods are trained and evaluated on the same datasets (KITTI, nuScenes, Waymo), which capture limited geographic, weather, and traffic diversity. Sim-to-real transfer techniques (domain randomization, adversarial domain adaptation, and digital twins) have shown promise for single-modality perception but remain largely unexplored for multi-modal fusion, where the joint distribution shift across sensors introduces additional complexity. Furthermore, long-tail scenarios

(rare object classes, unusual configurations, extreme weather events) are systematically underrepresented in benchmarks, yet these are precisely the conditions where robust fusion is most needed. The development of targeted long-tail evaluation protocols and fusion methods that explicitly model epistemic uncertainty for out-of-distribution inputs is an important open direction.

12.8. Standardized Benchmarks for Full-Spectrum Fusion

The community would benefit from new benchmark datasets that span the depth continuum: a dataset with the same scenes captured by 2D LiDAR, 16-beam, 32-beam, 64-beam, and 128-beam LiDAR simultaneously, alongside cameras with varying depth capabilities (mono, stereo, RGB-D), ideally across several application domains rather than driving alone (Section 11), with ground truth and metrics suited to each domain (e.g., poses for SLAM and map labels for semantic mapping). Such a dataset would enable systematic study of how fusion strategy and performance scale with depth information, providing the empirical foundation for the taxonomy proposed in this review, and in particular allowing the indicative 32-beam boundary for BEV fusion (Section 7.1) to be tested under controlled conditions.

13. Summary of Findings and Strength of Evidence

Table 12 summarizes the key findings of this review together with the strength of the evidence supporting each, following the three levels defined in Section 2. Findings that are consistently observed or hypothesized, rather than measured, mark where controlled experiments are still needed (Section 12).

Table 12. Key findings of the review and strength of the supporting evidence. Measured: controlled comparison or dedicated benchmark; consistently observed: recurring pattern across studies that are not directly comparable; hypothesized: architectural reasoning without direct validation.

Finding	Evidence Level	Basis and Limitations
(i) The fusion strategy should track the available depth	Consistently observed	Distribution of strategies across sensor tiers in the corpus (Table 2); trend in early vs. mid-level fusion across beam counts [44]; no controlled comparison of all strategies at all densities
(ii) Depth completion must be trained at the target LiDAR density	Measured	Cross-density evaluation of depth completion methods from 4 to 64 lines [73]
(iii) BEV fusion is documented from about 32 beams upward	Consistently observed; partially measured	Most BEV methods trained on 32-beam nuScenes; beam-reduction corruption (32→16/8/4) in MultiCorrupt [74], without retraining or depth-completion baseline (Section 7.1)

Table 12. *Cont.*

Finding	Evidence Level	Basis and Limitations
(iv) No standardized benchmark exists for 2D LiDAR–camera fusion	Measured (dataset survey)	Review of available datasets (Section 11)
(v) Late fusion keeps operating under single-sensor loss	Hypothesized; indirect evidence	Architectural reasoning; robustness of the decoupled SparseFusion in MultiCorrupt [74]; no benchmark includes a late-fusion detector (Section 10)
Calibration sensitivity decreases from hard projection to learned soft association	Measured	Two extrinsic-perturbation benchmarks [74,97] (Figure 6); late fusion not included

14. Conclusions

This review has presented a depth-centric perspective on LiDAR–camera fusion, arguing that the optimal fusion strategy is fundamentally determined by the quantity and quality of depth information available from the sensor configuration. Unlike existing surveys, which focus almost exclusively on multi-beam 3D LiDAR configurations, we have surveyed the full spectrum, from 2D LiDAR with depth cameras to 128-beam LiDAR systems, and proposed a three-axis taxonomy that maps sensor configurations to appropriate fusion strategies.

Several findings emerge from this survey. The first is that the fusion strategy should match the depth actually available: projection-based fusion suits 2D and low-beam LiDAR, depth completion bridges the sparse-to-dense gap, BEV fusion is documented from roughly 32 beams onward (an indicative lower bound rather than a controlled threshold), and attention-based fusion is the natural choice for dense 128+ beam configurations, although its reported results come from 32-beam data. Within this progression, depth completion stands out as the bridging mechanism that supplies low-cost LiDAR configurations with dense depth, although its effectiveness hinges on being trained and evaluated at the target LiDAR density. A second, more sobering finding is that the research community has yet to address the low-cost gap: the absence of standardized 2D LiDAR–camera benchmarks leaves academic work (focused on 32- to 64-beam sensors) disconnected from practical robotics (predominantly 2D LiDAR). Robustness, in turn, requires explicit modeling of sensor failure: late fusion provides graceful degradation under sensor loss by construction, whereas middle and BEV fusion need additional mechanisms, such as training with missing modalities, to achieve it; this contrast rests on architectural reasoning and indirect evidence, and still awaits direct benchmark validation. The strength of the evidence behind each of these findings is summarized in Section 13. Finally, the trajectory of the field is toward adaptivity: unified frameworks that dynamically adjust the fusion strategy according to the available depth, the environmental conditions, and the task at hand represent the next frontier.

By mapping sensor configurations to fusion strategies across the full depth continuum, this review provides a framework for navigating the rapidly expanding literature on LiDAR–camera fusion. We hope it proves useful both to researchers entering the field and to practitioners designing next-generation perception systems.

Author Contributions: Conceptualization, methodology, investigation, writing—original draft preparation, A.D.; review and editing, S.F., J.R., M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3DGS	3D Gaussian splatting
ADAS	Advanced driver-assistance system
AGV	Automated guided vehicle
ATE	Absolute trajectory error
BEV	Bird's eye view
CNN	Convolutional neural network
CSPN	Convolutional spatial propagation network
DC	Depth completion
fps	Frames per second
GPS	Global Positioning System
IMU	Inertial measurement unit
IoU	Intersection over union
IR	Infrared
iRMSE	Inverse root mean square error
ISO	International Organization for Standardization
LiDAR	Light Detection and Ranging
LSS	Lift–Splat–Shoot
MAE	Mean absolute error
mAP	Mean average precision
MDE	Monocular depth estimation
MEMS	Micro-electro-mechanical system
NDS	nuScenes Detection Score
NeRF	Neural radiance field
NIR	Near-infrared
PPS	Pulse-per-second
PTP	Precision Time Protocol
RA	Resistance ability
RGB	Red–green–blue
RGB-D	Red–green–blue plus depth
RMSE	Root mean square error
ROS	Robot Operating System
RPE	Relative pose error
SAE	Society of Automotive Engineers
SAM	Segment Anything Model
SLAM	Simultaneous localization and mapping
SOTIF	Safety of the intended functionality
ToF	Time-of-flight
UAMF	Uncertainty-aware multi-modal fusion
VLA	Vision–language–action

References

1. Tang, Y.; He, H.; Wang, Y.; Mao, Z.; Wang, H. Multi-modality 3D object detection in autonomous driving: A review. *Neurocomputing* **2023**, *553*, 126587. [CrossRef]
2. Wang, X.; Li, K.; Chehri, A. Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 1148–1165. [CrossRef]

3. Kumar, A.; Stephen, K.; Sabitha, A.S. A Systematic Review on Sensor Fusion Technology in Autonomous Vehicles. In Proceedings of the 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 6–8 July 2023; pp. 42–48. [CrossRef]
4. Wang, Y.; Mao, Q.; Zhu, H.; Deng, J.; Zhang, Y.; Ji, J.; Li, H.; Zhang, Y. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey. *Int. J. Comput. Vis.* **2023**, *131*, 2122–2152. [CrossRef]
5. Qian, H.; Wang, M.; Zhu, M.; Wang, H. A Review of Multi-Sensor Fusion in Autonomous Driving. *Sensors* **2025**, *25*, 6033. [CrossRef]
6. Valverde, M.; Moutinho, A.; Zacchi, J.V. A Survey of Deep Learning-Based 3D Object Detection Methods for Autonomous Driving Across Different Sensor Modalities. *Sensors* **2025**, *25*, 5264. [CrossRef]
7. Jiang, M.; Chu, Z.; Pan, Y.; Yuan, H.; Han, Q.; Zou, J.; Zhang, Y.; Zhu, X.; Zhang, C.; Zhou, X.; et al. Multi-sensor fusion on autonomous driving perception: A comprehensive survey. *Multimodal Transp.* **2026**, *5*, 100325. [CrossRef]
8. Yeong, D.J.; Velasco-Hernandez, G.; Barry, J.; Walsh, J. Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review. *Sensors* **2021**, *21*, 2140. [CrossRef]
9. Huang, K.; Shi, B.; Li, X.; Li, X.; Huang, S.; Li, Y. Multi-modal Sensor Fusion for Auto Driving Perception: A Survey. *arXiv* **2022**, arXiv:2202.02703.
10. Alaba, S.Y.; Gurbuz, A.C.; Ball, J.E. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. *World Electr. Veh. J.* **2024**, *15*, 20. [CrossRef]
11. Viktor, P.; Kiss, G. Multimodal Sensor Fusion in Autonomous Vehicles: Technologies, Architectures, and Open Challenges. *Sensors* **2026**, *26*, 3528. [CrossRef]
12. Wang, H.; Liu, J.; Dong, H.; Shao, Z. A Survey of the Multi-Sensor Fusion Object Detection Task in Autonomous Driving. *Sensors* **2025**, *25*, 2794. [CrossRef]
13. Wang, Y.; Wang, S.; Li, Y.; Liu, M. Developments in 3-D Object Detection for Autonomous Driving: A Review. *IEEE Sens. J.* **2025**, *25*, 21033–21053. [CrossRef]
14. Thakur, A.; Mishra, S.K. An in-depth evaluation of deep learning-enabled adaptive approaches for detecting obstacles using sensor-fused data in autonomous vehicles. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108550. [CrossRef]
15. Khatab, E.; Fathy, F.; AlKholy, A.; Shalash, O. Sensor Fusion and Perception for Autonomous Driving: A Critical Review of Modalities, AI Models, Algorithms, and Industry Configurations. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 199. [CrossRef]
16. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2012; pp. 3354–3361.
17. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 11618–11628.
18. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 2443–2451.
19. Iqbal, U.; Massoud, A.; Noureldin, A. A Sensor-Centric Survey of Autonomous Driving: Integrating Measurement Physics, Uncertainty Modeling, and Safety-Critical Multi-Sensor Fusion. *Sensors* **2026**, *26*, 3801. [CrossRef]
20. Viktor, P.; Kiss, G. Sensors in Self-Driving Vehicles: A Detailed Literature Review and New Trends. *Sensors* **2026**, *26*, 2153. [CrossRef]
21. Wang, L.; Zhang, X.; Song, Z.; Bi, J.; Zhang, G.; Wei, H.; Tang, L.; Yang, L.; Li, J.; Jia, C.; et al. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy. *IEEE Trans. Intell. Veh.* **2023**, *8*, 3781–3798. [CrossRef]
22. Zhong, H.; Wang, H.; Wu, Z.; Zhang, C.; Zheng, Y.; Tang, T. A survey of LiDAR and camera fusion enhancement. *Procedia Comput. Sci.* **2021**, *183*, 579–588. [CrossRef]
23. Al-Tawil, B.; Candemir, A.; Jung, M.; Al-Hamadi, A. Mobile robot navigation with enhanced 2D mapping and multi-sensor fusion. *Sensors* **2025**, *25*, 2408. [CrossRef]
24. Tatar, G.; Arar, M.E. Hardware-Accelerated 3D LiDAR-Based Object Detection with BEV Spatial Mapping on Embedded FPGA Platforms. *Electronics* **2026**, *15*, 2296. [CrossRef]
25. Liang, Y.; Hu, Y.; Shao, W.; Fu, Y. Distilling Monocular Foundation Model for Fine-grained Depth Completion. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025.
26. Li, A.; Hu, A.; Xi, W.; Yu, W.; Zou, D. Stereo-LiDAR Depth Estimation with Deformable Propagation and Learned Disparity-Depth Conversion. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 2729–2736. [CrossRef]
27. Meng, H.; Li, C.; Zhong, C.; Gu, J.; Chen, G.; Knoll, A. FastFusion: Deep stereo-LiDAR fusion for real-time high-precision dense depth sensing. *J. Field Robot.* **2023**, *40*, 1804–1816. [CrossRef]

28. Labbé, M.; Michaud, F. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *J. Field Robot.* **2019**, *36*, 416–446. [CrossRef]
29. Messbah, H.; Emharraf, M.; Saber, M.; Al Kaddouri, H. Sensor Fusion Using LiDAR and Visual Sensor for SLAM Indoor Navigation. In *Recent Advances in Robotics and Automation (Lecture Notes in Electrical Engineering)*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 321–331.
30. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; Dai, J. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
31. Zhao, J.; Shi, J.; Zhuo, L. BEV perception for autonomous driving: State of the art and future perspectives. *Expert Syst. Appl.* **2024**, *258*, 125103. [CrossRef]
32. Ma, Y.; Wang, T.; Bai, X.; Yang, H.; Hou, Y.; Wang, Y.; Qiao, Y.; Yang, R.; Manocha, D.; Zhu, X. Vision-Centric BEV Perception: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10978–10997. [CrossRef]
33. Obando-Ceron, J.S.; Romero-Cano, V.; Monteiro, S. Probabilistic multi-modal depth estimation based on camera–LiDAR sensor fusion. *Mach. Vis. Appl.* **2023**, *34*, 79. [CrossRef]
34. Nie, X.; Zhu, L.; He, Z.; Cheng, A.; Zhong, S.; Li, E. Investigating 3D object detection using stereo camera and LiDAR fusion with bird's-eye view representation. *Neurocomputing* **2025**, *620*, 129144. [CrossRef]
35. Liu, H.; Wu, C.; Wang, H. Real time object detection using LiDAR and camera fusion for autonomous driving. *Sci. Rep.* **2023**, *13*, 8056. [CrossRef]
36. Vinoth, K.; Sasikumar, P. Multi-sensor fusion and segmentation for autonomous vehicle multi-object tracking using deep Q networks. *Sci. Rep.* **2024**, *14*, 31130. [CrossRef]
37. Li, Z.; Singh, B.S.B. Autonomous driving in adverse weather: A multi-modal fusion framework with uncertainty-aware learning for robust obstacle detection. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*. [CrossRef]
38. Palladin, E.; Dietze, R.; Narayanan, P.; Bijelic, M.; Heide, F. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 484–503.
39. Zhang, B.; Wang, Y.; Zhang, C.; Jiang, J.; Luo, X.; Wang, X.; Zhang, Y.; Liu, Z.; Shen, G.; Ye, Y.; et al. FogFusion: Robust 3D object detection based on camera-LiDAR fusion for autonomous driving in foggy weather conditions. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2026**, *240*, 2312–2322. [CrossRef]
40. Cheng, L.; Guo, L.; Zhang, T.; Bang, T.; Harris, A.; Hajij, M.; Sartipi, M.; Cao, S. CalibRefine: Deep Learning-Based Online Automatic Targetless LiDAR–Camera Calibration with Iterative and Attention-Driven Post-Refinement. *IEEE Trans. Instrum. Meas.* **2026**, *75*, 1–18. [CrossRef]
41. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 4603–4611.
42. Wang, C.; Ma, C.; Zhu, M.; Yang, X. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 11789–11798.
43. Benayed, W.; Mabrouk, I.; Masmoudi, M.S.; Abdelaziz, W.B. LiDAR 2D and Camera Fusion for ADAS: A Practical Approach With YOLOv9 and ROS2 Framework. *IEEE Access* **2025**, *13*, 155500–155519. [CrossRef]
44. Mousa-Pasandi, M.; Liu, T.; Massoud, Y.; Laganière, R. RGB-LiDAR fusion for accurate 2D and 3D object detection. *Mach. Vis. Appl.* **2023**, *34*, 86. [CrossRef]
45. Gan, X.; Zhu, K.; Sun, M.; Zhao, L.; Lai, C. GAC-Net: A Geometric–Attention Fusion Network for Sparse Depth Completion from LiDAR and Image. *Sensors* **2025**, *25*, 5495. [CrossRef]
46. Wei, M.; Zhu, M.; Zhang, Y.; Wang, J.; Sun, J. Real-time depth completion based on LiDAR-stereo for autonomous driving. *Front. Neurorobotics* **2023**, *17*, 1124676. [CrossRef]
47. Jiao, Y.; Jie, Z.; Chen, S.; Chen, J.; Ma, L.; Jiang, Y.G. MSMDFusion: Fusing LiDAR and Camera at Multiple Scales with Multi-Depth Seeds for 3D Object Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
48. Yu, X.; Lu, K.; Yang, Y.; Ou, L. DMFusion: LiDAR-camera fusion framework with depth merging and temporal aggregation. *Appl. Intell.* **2024**, *54*, 9412–9428. [CrossRef]
49. Cai, H.; Zhang, Z.; Zhou, Z.; Li, Z.; Ding, W.; Zhao, J. BEVFusion4D: Learning LiDAR-Camera Fusion Under Bird's-Eye-View via Cross-Modality Guidance and Temporal Aggregation. *arXiv* **2023**, arXiv:2303.17099.
50. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2023; pp. 2774–2781.

51. Li, X.; Fan, B.; Tian, J.; Fan, H. Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 21209–21218.
52. Yang, Z.; Chen, J.; Miao, Z.; Li, W.; Zhu, X.; Zhang, L. Deepinteraction: 3D object detection via modality interaction. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 1992–2005.
53. Massoud, Y.; Laganiere, R. Learnable fusion mechanisms for multimodal object detection in autonomous vehicles. *IET Comput. Vis.* **2024**, *18*, 499–511. [CrossRef]
54. Tran, N.A.; Hong, Y.G. Adaptive Fusion of LiDAR and Camera Data for Enhanced Precision in 3D Object Detection for Autonomous Driving. *IEEE Access* **2025**, *13*, 155474–155481. [CrossRef]
55. Qiao, R.; Yuan, H.; Guan, Z.; Zhang, W. MDFusion: Multi-Dimension Semantic–Spatial Feature Fusion for LiDAR–Camera 3D Object Detection. *Remote Sens.* **2025**, *17*, 1240. [CrossRef]
56. Hu, X.; Duan, Z.; An, P.; Zhang, J.; Ma, J. Lidar-camera range-view fusion for 3D object detection in autonomous driving. *Multimed. Syst.* **2025**, *31*, 363. [CrossRef]
57. He, Y.; Zhao, L.; Deng, T.; Fang, Z.; Chen, W. An Efficient LiDAR-Camera Fusion Network for Multiclass 3-D Dynamic Object Detection and Trajectory Prediction. *IEEE Trans. Instrum. Meas.* **2026**, *75*, 2510615. [CrossRef]
58. Yu, L.; Zhang, J.; Liu, Z.; Yue, H.; Chen, W. Rethinking the late fusion of lidar-camera based 3d object detection. In *Proceedings of the 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
59. Ali, A.; Tawfik, M.M.; Saafan, M.M. Cross-dataset late fusion of Camera–LiDAR and radar models for object detection. *Sci. Rep.* **2026**, *16*, 879. [CrossRef]
60. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 12689–12697.
61. Dai, Y.; Kim, D.; Lee, K. An Advanced Approach to Object Detection and Tracking in Robotics and Autonomous Vehicles Using YOLOv8 and LiDAR Data Fusion. *Electronics* **2024**, *13*, 2250. [CrossRef]
62. Fan, Z.; Zhang, L.; Wang, X.; Shen, Y.; Deng, F. LiDAR, IMU, and camera fusion for simultaneous localization and mapping: A systematic review. *Artif. Intell. Rev.* **2025**, *58*, 174. [CrossRef]
63. Zhu, J.; Li, H.; Zhang, T. Camera, LiDAR, and IMU based multi-sensor fusion SLAM: A survey. *Tsinghua Sci. Technol.* **2023**, *29*, 415–429. [CrossRef]
64. Lv, X.; He, Z.; Yang, Y.; Nie, J.; Dong, Z.; Wang, S.; Gao, M. MSF-SLAM: Multi-sensor-fusion-based simultaneous localization and mapping for complex dynamic environments. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 19699–19713. [CrossRef]
65. Yue, X.; Zhang, Y.; Chen, J.; Chen, J.; Zhou, X.; He, M. LiDAR-based SLAM for robotic mapping: State of the art and new frontiers. *Ind. Robot. Int. J. Robot. Res. Appl.* **2024**, *51*, 196–205. [CrossRef]
66. Dai, Z.; Guan, Z.; Chen, Q.; Xu, Y.; Sun, F. Enhanced Object Detection in Autonomous Vehicles through LiDAR—Camera Sensor Fusion. *World Electr. Veh. J.* **2024**, *15*, 297. [CrossRef]
67. Liu, C.; Zhang, G.; Rong, Y.; Shao, W.; Meng, J.; Li, G.; Huang, Y. Hybrid metric-feature mapping based on camera and Lidar sensor fusion. *Measurement* **2023**, *207*, 112411. [CrossRef]
68. Xia, Y.; Wu, H.; Zhu, L.; Qi, W.; Zhang, S.; Zhu, J. A multi-sensor fusion framework with tight coupling for precise positioning and optimization. *Signal Process.* **2024**, *217*, 109343. [CrossRef]
69. Tian, Y.; Zhang, J.; Li, L.; Zhang, D.; Zuo, Y.; Yin, J. Ultra-Fusion: A Resilient Tightly-Coupled Multi-Sensor Fusion SLAM Framework under Sensor Degradation and Spatiotemporal Perturbation for Intelligent Transportation Systems. *arXiv* **2026**, arXiv:2606.21223.
70. Chen, H.; Ma, Z.; Li, Z.; Wang, P.; Xiao, Y. Mobile Robot SLAM Based on Multi-Sensor Fusion. In *Proceedings of the 2025 5th International Conference on Sensors and Information Technology*; IEEE: New York, NY, USA, 2025; pp. 777–780.
71. Gou, G.; Li, H.; Wang, X.; Zhang, H.; Yang, W.; Sui, H. Unsupervised deep depth completion with heterogeneous LiDAR and RGB-D camera depth information. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104327. [CrossRef]
72. Park, H.; Jeong, J.; Yoon, K.J. RAF: Reliability-Aware Fusion of Camera, LiDAR, and 4D RADAR for Robust 3D Object Detection in Adverse Weather. In *Proceedings of the Computer Vision—ECCV 2026*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2026; pp. 201–220. [CrossRef]
73. Park, H.; Gupta, A.; Wong, A. Test-Time Adaptation for Depth Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
74. Beemelmanns, T.; Zhang, Q.; Geller, C.; Eckstein, L. Multicorrupt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection. In Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Republic of Korea, 2–5 June 2024; pp. 3255–3261.

75. Li, H.; Zhao, Y.; Zhong, J.; Wang, B.; Sun, C.; Sun, F. Delving Into the Secrets of BEV 3D Object Detection in Autonomous Driving: A Comprehensive Survey. *IEEE Trans. Intell. Transp. Syst.* **2026**, *27*, 119–144. [CrossRef]
76. Gunn, J.; Lenyk, Z.; Sharma, A.; Donati, A.; Buburuzan, A.; Redford, J.; Mueller, R. Lift-attend-splat: Bird's-eye-view camera-lidar fusion using transformers. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2024; pp. 4526–4536.
77. Oladele, D.A.; Markus, E.D.; Abu-Mahfouz, A.M. BEV-CAM3D: A Unified Bird's-Eye View Architecture for Autonomous Driving with Monocular Cameras and 3D Point Clouds. *AI* **2025**, *6*, 82. [CrossRef]
78. Yu, Z.; Li, J.; Wei, Y.; Lyu, Y.; Tan, X. Combining Camera–LiDAR Fusion and Motion Planning Using Bird's-Eye View Representation for End-to-End Autonomous Driving. *Drones* **2025**, *9*, 281. [CrossRef]
79. Harley, A.W.; Fang, Z.; Li, J.; Ambrus, R.; Fragkiadaki, K. Simple-BEV: What Really Matters for Multi-Sensor BEV Perception? In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023.
80. Zhao, Y.; Gong, Z.; Zheng, P.; Zhu, H.; Wu, S. SimpleBEV: Improved LiDAR-Camera Fusion Architecture for 3D Object Detection. *arXiv* **2024**, arXiv:2411.05292.
81. Shi, P.; Liu, Z.; Dong, X.; Yang, A. CL-fusionBEV: 3D object detection method with camera-LiDAR fusion in Bird's Eye View. *Complex Intell. Syst.* **2024**, *10*, 7681–7696. [CrossRef]
82. Shi, P.; Yang, L.; Dong, X.; Qi, H.; Yang, A. Research progress on multi-modal fusion object detection algorithms for autonomous driving: A review. *Comput. Mater. Contin.* **2025**, *83*, 3877–3917. [CrossRef]
83. Yin, C.; Qu, H.; Li, Y. 3D Object Detection Algorithm for Autonomous Driving Based on Multi-Scale Feature Weighted Point-by-Point Fusion. *J. Intell. Fuzzy Syst.* **2025**, *49*, 933–947. [CrossRef]
84. Ma, F.; Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2018; pp. 4796–4803.
85. Cheng, X.; Wang, P.; Yang, R. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 108–125.
86. Fan, Y.; Chen, Z.; Liu, H.; Xu, K. Holistic and Contextual Evidential Stereo-LiDAR Fusion for Depth Estimation. *IEEE Trans. Intell. Veh.* **2024**, *9*, 7437–7448. [CrossRef]
87. Shuo, W.; Zhang, P.; Yihong, W. Review of extrinsic parameter calibration of LiDAR and camera. *Virtual Real. Intell. Hardw.* **2026**, *8*, 28–70. [CrossRef]
88. Qiu, Z.; Martínez-Sánchez, J.; Arias-Sanchez, P.; Rashdi, R. External multi-modal imaging sensor calibration for sensor fusion: A review. *Inf. Fusion* **2023**, *97*, 101806. [CrossRef]
89. Tan, Z.; Zhang, X.; Teng, S.; Wang, L.; Gao, F. A review of deep learning-based LiDAR and camera extrinsic calibration. *Sensors* **2024**, *24*, 3878. [CrossRef]
90. An, P.; Ding, J.; Quan, S.; Yang, J.; Yang, Y.; Liu, Q.; Ma, J. Survey of extrinsic calibration on lidar-camera system for intelligent vehicle: Challenges, approaches, and trends. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 15342–15366. [CrossRef]
91. Cipelli, L.; D'Addeo, F.; Ghelfi, E.; Ceresini, M.; Bertogalli, A.; Pirazzoli, F.; Bertozzi, M. End-to-End LiDAR-Camera Calibration via Multi-Modal Correspondences Estimation and Explicit BEV Alignment. *Int. J. Comput. Vis.* **2026**, *134*, 333. [CrossRef]
92. Yan, G.; He, F.; Shi, C.; Wei, P.; Cai, X.; Li, Y. Joint camera intrinsic and LiDAR-camera extrinsic calibration. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 11446–11452.
93. Wang, C.; Liu, S.; Wang, X.; Lan, X. Time Synchronization and Space Registration of Roadside LiDAR and Camera. *Electronics* **2023**, *12*, 537. [CrossRef]
94. Song, Z.; Peng, L.; Hu, J.; Yao, D.; Zhang, Y. Timealign: A multi-modal object detection method for time misalignment fusing in autonomous driving. *arXiv* **2024**, arXiv:2412.10033.
95. Lu, Y.; Zhong, W.; Li, Y. Calibration of multi-sensor fusion for autonomous vehicle system. *Int. J. Veh. Des.* **2023**, *91*, 248. [CrossRef]
96. Gurumadaiah, A.K.; Park, J.; Lee, J.; Kim, J.; Kwon, S. Precise Synchronization Between LiDAR and Multiple Cameras for Autonomous Driving: An Adaptive Approach. *IEEE Trans. Intell. Veh.* **2025**, *10*, 2152–2162. [CrossRef]
97. Yu, K.; Tao, T.; Xie, H.; Lin, Z.; Liang, T.; Wang, B.; Chen, P.; Hao, D.; Wang, Y.; Liang, X. Benchmarking the robustness of lidar-camera fusion for 3d object detection. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE: New York, NY, USA, 2023; pp. 3188–3198.
98. Gu, Z.; Wu, Z.; Zhang, C.; Liang, Y. Point cloud processing under adverse weather: A survey of datasets, enhancement, and denoising methods. *Signal Image Video Process.* **2025**, *19*, 718. [CrossRef]
99. Yang, X.; Wang, L. AWAD-Fusion: Dynamic Multi-sensor Fusion Framework for Robust 3D Object Detection in Adverse Weather. In *Pattern Recognition and Computer Vision (Lecture Notes in Computer Science)*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 132–143.
100. Gupta, H.; Lilienthal, A.J.; Andreasson, H. Evaluating lidar perception algorithms for all-weather autonomy. *Sensors* **2025**, *25*, 7436. [CrossRef]

101. Feng, S.; Cai, X.; Li, L.; Wang, W.; Ying, S. A review of research on vehicle detection in adverse weather environments. *J. Traffic Transp. Eng. (Engl. Ed.)* **2025**, *12*, 1452–1483. [CrossRef]
102. Thottempudi, P.; Jambek, A.B.B.; Kumar, V.; Acharya, B.; Moreira, F. Resilient object detection for autonomous vehicles: Integrating deep learning and sensor fusion in adverse conditions. *Eng. Appl. Artif. Intell.* **2025**, *151*, 110563. [CrossRef]
103. Alam, M.Z.; Kaleem, Z.; Kelouwani, S. Glare Mitigation for Enhanced Autonomous Vehicle Perception. *IEEE Trans. Intell. Veh.* **2024**, *9*, 7030–7044. [CrossRef]
104. Sun, Y.; Dai, C.; Li, W.; Liu, X.; Sun, Y.; Zhang, Y.; Guan, W.; Zhang, Y.; Guo, Y.; Wang, H. RTPSeg: A multi-modality dataset for LiDAR point cloud semantic segmentation assisted with RGB-thermal images in autonomous driving. *ISPRS J. Photogramm. Remote Sens.* **2026**, *233*, 25–38. [CrossRef]
105. ROVR Network. ROVR Open Dataset. 2025. Available online: https://github.com/rovr-network/ROVR-Open-Dataset (accessed on 21 October 2026).
106. Qian, J.; Chatrath, V.; Yang, J.; Servos, J.; Schoellig, A.P.; Waslander, S.L. POCD: Probabilistic Object-Level Change Detection and Volumetric Mapping in Semi-Static Scenes. In Proceedings of the Robotics: Science and Systems XVIII, New York, NY, USA, 27 June–1 July 2022. [CrossRef]
107. Rüdt, M.; Pang, H.; Enke, C.; Furmans, K.; Seibold, Z. *Contextual Semantic Mapping Dataset for Intralogistics: RGB Images, 2D LiDAR Scans, and SAM Instance Annotations*; RADAR4KIT; Karlsruhe Institute of Technology: Karlsruhe, Germany, 2026. [CrossRef]
108. Chebrolu, N.; Lottes, P.; Schaefer, A.; Winterhalter, W.; Burgard, W.; Stachniss, C. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Robot. Res.* **2017**, *36*, 1045–1052. [CrossRef]
109. Ruiz-Sarmiento, J.R.; Galindo, C.; Gonzalez-Jimenez, J. Robot@Home, a robotic dataset for semantic mapping of home environments. *Int. J. Robot. Res.* **2017**, *36*, 131–141. [CrossRef]
110. Shi, X.; Li, D.; Zhao, P.; Tian, Q.; Tian, Y.; Long, Q.; Zhu, C.; Song, J.; Qiao, F.; Song, L.; et al. Are We Ready for Service Robots? The OpenLORIS-Scene Datasets for Lifelong SLAM. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 3139–3145. [CrossRef]
111. Martín-Martín, R.; Patel, M.; Rezatofighi, H.; Shenoi, A.; Gwak, J.; Frankel, E.; Sadeghian, A.; Savarese, S. JRDB: A Dataset and Benchmark of Egocentric Robot Visual Perception of Humans in Built Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 6748–6765. [CrossRef]
112. Fallon, M.; Johannsson, H.; Kaess, M.; Leonard, J.J. The MIT Stata Center dataset. *Int. J. Robot. Res.* **2013**, *32*, 1695–1699. [CrossRef]
113. Jia, D.; Steinweg, M.; Hermans, A.; Leibe, B. Self-Supervised Person Detection in 2D Range Data using a Calibrated Camera. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13301–13307. [CrossRef]
114. Xie, Z.; Pan, Y.; Zhang, Y.; Pan, J.; Dames, P. Semantic2D: Enabling Semantic Scene Understanding with 2D Lidar Alone. *arXiv* **2024**, arXiv:2409.09899.
115. Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 573–580. [CrossRef]
116. Lin, C.; Sun, W.; Sun, G.; Gong, B.; Liu, H. Onboard LiDAR–camera deployment optimization for pavement marking distress fusion detection. *Sensors* **2025**, *25*, 3875. [CrossRef]
117. Fourati, S.; Jaafar, W.; Baccar, N.; Alfattani, S.; Langar, R. Foundation models for autonomous driving: A comprehensive survey. *Eng. Appl. Artif. Intell.* **2026**, *176*, 114805. [CrossRef]
118. Din, M.U.; Akram, W.; Saoud, L.S.; Rosell, J.; Hussain, I. Multimodal fusion with vision-language-action models for robotic manipulation: A systematic review. *Inf. Fusion* **2026**, *129*, 104062. [CrossRef]
119. Deng, K.; Xing, L.; Wu, H.; Ma, H.; Ling, Y.; Gao, J. Advances in object detection for autonomous driving using mmwave radar and camera: A comprehensive survey. *J. King Saud Univ. Comput. Inf. Sci.* **2025**, *37*, 328. [CrossRef]
120. Xu, H.; Chen, J.; Meng, S.; Wang, Y.; Chau, L.P. A survey on occupancy perception for autonomous driving: The information fusion perspective. *Inf. Fusion* **2025**, *114*, 102671. [CrossRef]
121. *ISO 21448:2022*; Road Vehicles—Safety of the Intended Functionality. International Organization for Standardization: Geneva, Switzerland, 2022.
122. *ISO 26262:2018*; Road Vehicles—Functional Safety. International Organization for Standardization: Geneva, Switzerland, 2018.
123. Yeong, D.J.; Panduru, K.; Walsh, J. Exploring the unseen: A survey of multi-sensor fusion and the role of explainable ai (xai) in autonomous vehicles. *Sensors* **2025**, *25*, 856. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.