Manuscript Template for Opto-Electronic Journals

A Survey of LiDAR–Camera Calibration: From Explicit Geometric Constraints to Differentiable Scene Representations

Fengli Yang[1], Jiangnan Peng[1], Zhihong Zeng[1], Dengke Wang[1], Chen Chen*[1], Long Zhao*[2], and Harald Haas[3]

[1]**School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China**

[2]**School of Space and Earth Sciences, Beihang University, Beijing 100191, China**

[3]**Department of Engineering, Electrical Engineering Division, Cambridge University, CB3 0FA Cambridge, UK**

Abstract: LiDAR–camera extrinsic calibration is essential for reliable multi-sensor perception, as even small pose errors can induce noticeable misalignment between projected point clouds and image structures. While the problem reduces to estimating a six-degree-of-freedom rigid transformation, its fundamental challenge lies in establishing reliable constraints across heterogeneous LiDAR and camera measurements. This survey organizes the literature along the central theme of "from explicit geometric constraints to differentiable scene representations". Existing methods fall into four categories: target-based calibration using explicit geometric constraints, targetless calibration via natural-scene matching, learning-based cross-modal alignment with online correction, and joint calibration leveraging differentiable scene representations. For each category, we examine typical evidence sources, representative optimization formulations, advantages, and limitations. Special focus is placed on recent advances in NeRF, 3D Gaussian Splatting, and 2D Gaussian Splatting, which reformulate calibration from local correspondence estimation into a scene-level differentiable optimization problem. The survey concludes by identifying an emerging trend toward hybrid systems that synergistically integrate explicit geometric cues, learned priors, uncertainty quantification, and efficient differentiable rendering, thereby enabling robust online multi-sensor calibration.

Keywords: LiDAR–camera calibration; extrinsic calibration; cross-modal alignment; targetless calibration; differentiable scene representations; Gaussian Splatting

1 Introduction

LiDAR and cameras are among the most widely adopted heterogeneous sensor pairs in mobile robotics, autonomous driving, three-dimensional reconstruction, and augmented reality. LiDAR provides metric three-dimensional

measurements that describe scene geometry, spatial layout, and object contours with high geometric fidelity. Cameras, in contrast, capture dense color, texture, and semantic information that supports object recognition, scene understanding, and fine-grained visual analysis. The complementarity between metric geometry and visual appearance makes image–point cloud fusion a cornerstone of modern intelligent perception systems.

Such fusion is reliable only when the two sensors are accurately related within a common spatial reference frame. LiDAR–camera extrinsic calibration estimates the relative pose between the LiDAR and camera coordinate systems, enabling LiDAR points to be projected consistently onto the image plane and allowing image-derived texture or semantic information to be associated with three-dimensional structures. Errors in the extrinsic parameters cause spatial misalignment between images and point clouds, particularly near object boundaries, image edges, and depth discontinuities. These errors can propagate to downstream tasks such as object detection, semantic segmentation, three-dimensional reconstruction, localization, and multi-sensor fusion. Consequently, accurate, stable, and reproducible extrinsic calibration is a prerequisite for reliable LiDAR–camera perception.

Although extrinsic calibration reduces to estimating a six-degree-of-freedom rigid transformation, the main difficulty lies not in the numerical solution of rotation and translation per se, but in how to establish meaningful constraints across the two modalities. Cameras observe perspective-projected two-dimensional intensity, color, and texture patterns, whereas LiDAR measures sparse three-dimensional points, ranges, and reflectance intensities. The two sensors differ in imaging mechanism, dimensionality, sampling density, and feature representation, and they do not provide natural one-to-one correspondences. The central problem, therefore, is how to construct reliable cross-modal constraints: what evidence can be observed by both sensors, how to represent that evidence, and how to infer the extrinsic parameters from it.

Given the breadth of this problem, several recent surveys have emerged. Existing reviews have examined automatic targetless calibration, external calibration of multi-modal imaging sensors, LiDAR–camera calibration for intelligent vehicles, learning-based extrinsic calibration, calibration of 3-D LiDAR and monocular cameras, and general LiDAR–camera extrinsic parameter calibration [1–6]. These studies provide valuable taxonomies, technical summaries, and application-oriented discussions. However, most existing surveys organize the field primarily by calibration target, sensor configuration, or algorithmic category. In contrast, this review focuses on how cross-modal evidence is constructed, represented, and optimized. This perspective is useful because it connects seemingly different methods through a common question: how does each method turn heterogeneous observations into constraints on the extrinsic parameters.

Existing methods can be broadly interpreted as different answers to this question. Target-based methods introduce controllable calibration objects—such as checkerboards, planar boards, circular-hole boards, spherical targets, and reflective patterns—to construct explicit geometric constraints. Targetless methods eliminate the

need for dedicated calibration objects and instead exploit structures available in natural scenes, including edges, planes, line features, semantic objects, and statistical correlations [7–11]. Learning-based methods use neural networks to learn cross-modal correspondences or directly estimate extrinsic deviations from image and point-cloud data, thereby improving automation and online adaptability [4, 12–16]. More recently, differentiable scene representation methods have incorporated extrinsic parameters, scene representations, and rendering consistency into a unified optimization framework, shifting calibration from local correspondence estimation toward joint scene-level interpretation [17–24]. This methodological evolution is summarized in Fig. 1.

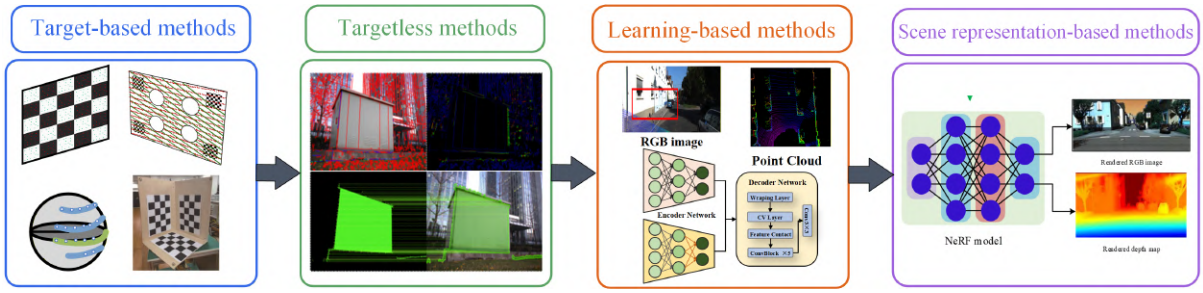

Figure 1: Technical evolution of LiDAR–camera extrinsic calibration methods.

Table 1 summarizes representative surveys and highlights the positioning of this review. The comparison shows that existing reviews have provided valuable taxonomies from the perspectives of targetless calibration, multi-modal sensor calibration, intelligent-vehicle systems, learning-based methods, and general LiDAR–camera calibration. In contrast, this review focuses on how cross-modal evidence is constructed, represented, and optimized.

Table 1: Representative surveys related to LiDAR–camera calibration and the positioning of this review.

Survey	**Main focus**	**Position of this review**
Li et al. [1]	Targetless calibration	Extends targetless explicit matching to a broader spectrum including target-based, learning-based, and differentiable scene representation methods.
Liu et al. [2]	Multi-modal imaging sensor calibration	Focuses specifically on LiDAR–camera calibration and emphasizes evidence construction and scene representation.
An et al. [3]	Intelligent-vehicle LiDAR–camera systems	Moves beyond correspondence categories to discuss how shared evidence is represented and optimized.
Tan et al. [4]	Deep learning-based calibration	Treats learning-based calibration as a transition from explicit geometry to differentiable scene-level optimization.
Zhang et al. [5]	3-D LiDAR and monocular camera calibration	Emphasizes the evolution of calibration formulations and recent differentiable rendering-based methods.
Virtual Reality & Intelligent Hardware [6]	LiDAR–camera extrinsic calibration review	Treats differentiable scene representations as an independent frontier rather than a supplementary trend.

Based on this positioning, this review makes three main contributions. First, it reconstructs the methodological landscape of LiDAR–camera calibration by examining the interplay among shared observational evidence, scene representation, and extrinsic parameter estimation. Second, it unifies target-based calibration, targetless explicit matching, learning-based cross-modal alignment, and differentiable scene representation methods within a coherent narrative, thereby highlighting the continuity across these research lines. Third, it analyzes how NeRF, 3D Gaussian Splatting, 2D Gaussian Splatting, and related differentiable scene representations reshape the formulation of extrinsic calibration, providing a basis for analyzing recent frontier methods.

Accordingly, this review characterizes the evolution of LiDAR–camera extrinsic calibration as a progression: from manually constructed geometric evidence, through automatically discovered structural evidence in natural scenes, to learned implicit correspondences, and finally to differentiable joint optimization under unified scene representations. Following this logic, Section 2 introduces the problem definition and basic mathematical model. Sections 3–6 then review, respectively, target-based explicit geometric constraints, targetless explicit matching in natural scenes, learning-based cross-modal alignment and online calibration, and joint calibration based on differentiable scene representations.

2 Problem Definition and Basic Model

As illustrated in Fig. 2, LiDAR–camera extrinsic calibration seeks to determine the rigid transformation between the LiDAR coordinate system and the camera coordinate system. Let $O_L - X_L Y_L Z_L$ denote the LiDAR frame and $O_C - X_C Y_C Z_C$ the camera frame. The extrinsic transformation from the LiDAR frame to the camera frame is described by the rotation matrix $\boldsymbol{R}_L^C$ and the translation vector $\boldsymbol{t}_L^C$. Given an image observation set $\mathcal{I} = \{\boldsymbol{p}_i\}_{i=1}^K$ and a LiDAR point cloud $\mathcal{P} = \{\boldsymbol{P}_i^L\}_{i=1}^N$, the objective is to estimate $\boldsymbol{R}_L^C$ and $\boldsymbol{t}_L^C$ such that the transformed LiDAR points are consistent with the corresponding image evidence under the camera projection model.

Under the standard pinhole camera model, a LiDAR point $\boldsymbol{P}_i^L = [X_i^L, Y_i^L, Z_i^L]^\mathsf{T}$ is first transformed into the camera coordinate frame as

$$\boldsymbol{P}_i^C = \boldsymbol{R}_L^C \boldsymbol{P}_i^L + \boldsymbol{t}_L^C \tag{1}$$

where $\boldsymbol{P}_i^C = [X_i^C, Y_i^C, Z_i^C]^\mathsf{T}$ denotes the same three-dimensional point expressed in the camera coordinate frame. In homogeneous coordinates, the LiDAR–camera extrinsic transformation is written as

$$\boldsymbol{T}_L^C = \begin{bmatrix} \boldsymbol{R}_L^C & \boldsymbol{t}_L^C \\ \boldsymbol{0}_{3\times1}^\mathsf{T} & 1 \end{bmatrix} \in SE(3) \tag{2}$$

The projection relationship between the 3D LiDAR point and its image point $\boldsymbol{p}_i = [u_i, v_i]^\mathsf{T}$ can be expressed as

$$Z_i^C \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \boldsymbol{K} \begin{bmatrix} \boldsymbol{R}_L^C & \boldsymbol{t}_L^C \end{bmatrix} \begin{bmatrix} X_i^L \\ Y_i^L \\ Z_i^L \\ 1 \end{bmatrix}, \tag{3}$$

where $\boldsymbol{K}$ is the camera intrinsic matrix, which is usually estimated in advance.

Equation (3) shows that extrinsic calibration directly governs the spatial alignment between LiDAR point

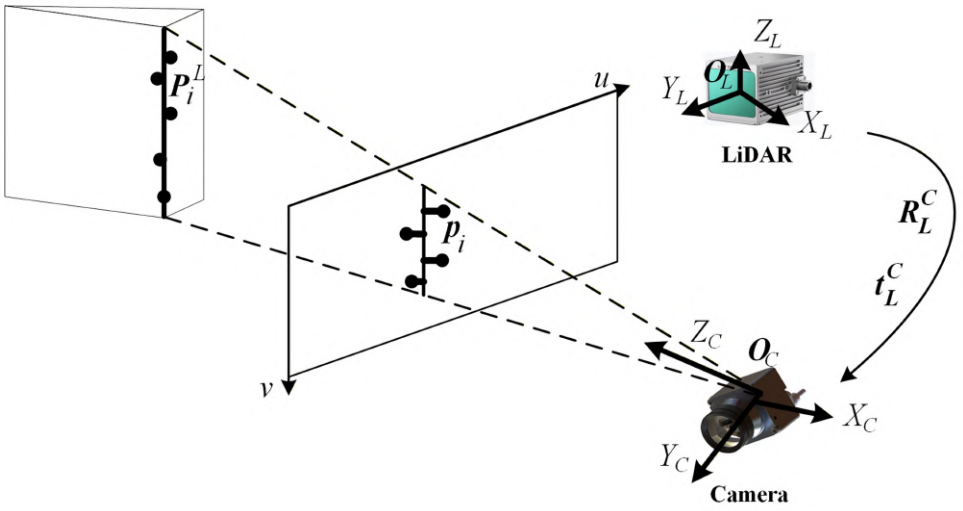

Figure 2: Geometric illustration of projecting a LiDAR point $\boldsymbol{P}_i^L$ onto the image plane.

clouds and camera images. If either $\boldsymbol{R}_L^C$ or $\boldsymbol{t}_L^C$ is inaccurate, the projected image point $\boldsymbol{p}_i$ will deviate from the corresponding image structure, thereby breaking the consistency among image edges, object contours, semantic regions, and depth boundaries. Consequently, most LiDAR–camera calibration methods can be interpreted as searching for the extrinsic transformation $\boldsymbol{T}_L^C$ that best aligns cross-modal observations under a specific type of constraint—be it geometric, semantic, statistical, learning-based, or rendering-based.

From an optimization perspective, although different methods use different observations, they can usually be abstracted as the minimization of a set of residual terms:

$$(\boldsymbol{T}_L^C)^* = \arg\min_{\boldsymbol{T}_L^C} \sum_i \rho\left(e_i(\boldsymbol{T}_L^C)\right), \tag{4}$$

where $e_i(\boldsymbol{T}_L^C)$ denotes the i-th cross-modal residual constructed under a candidate extrinsic transformation, and $\rho(\cdot)$ is a cost function or robust kernel used to reduce the influence of outliers. The specific form of $e_i(\boldsymbol{T}_L^C)$ depends on the observational cues employed by a given method. In target-based methods, the residual may be constructed from correspondences of corners, circle centers, planes, or sphere centers. In targetless methods, it may be derived from edges, line structures, planes, semantic regions, or statistical correlations. In learning-based methods, it may arise from predicted feature correspondences, extrinsic deviations, or geometric consistency losses. In differentiable scene representation methods, it can be further defined by discrepancies among rendered images, rendered depth maps, scene representations, and raw sensor observations.

3 Target-Based Methods: Explicit Geometric Constraints

Target-based calibration is one of the most established paradigms for LiDAR–camera extrinsic calibration. Its core idea is to introduce a structured object that can be reliably observed by both sensors. Rather than searching for correspondences in an unconstrained scene, these methods actively create shared evidence through known geometry, topology, texture, or reflectance. The calibration problem is thus converted from ambiguous cross-modal matching into extrinsic estimation under controlled geometric constraints.

This design is motivated by the substantial modality gap between LiDAR and cameras. LiDAR provides sparse metric three-dimensional measurements, whereas cameras deliver dense two-dimensional appearance information. The two modalities differ in sampling density, imaging mechanism, and feature representation. Artificial targets narrow this gap by forcing both sensors to observe the same physical structure. On the LiDAR side, planes, boundaries, hole contours, sphere centers, or intensity clusters can be extracted. On the camera side, corners, edges, markers, circles, ellipses, or projected centers can be detected. These features then instantiate the residual term $e_i(\boldsymbol{T}_L^C)$ in the general optimization model introduced in Section 2.

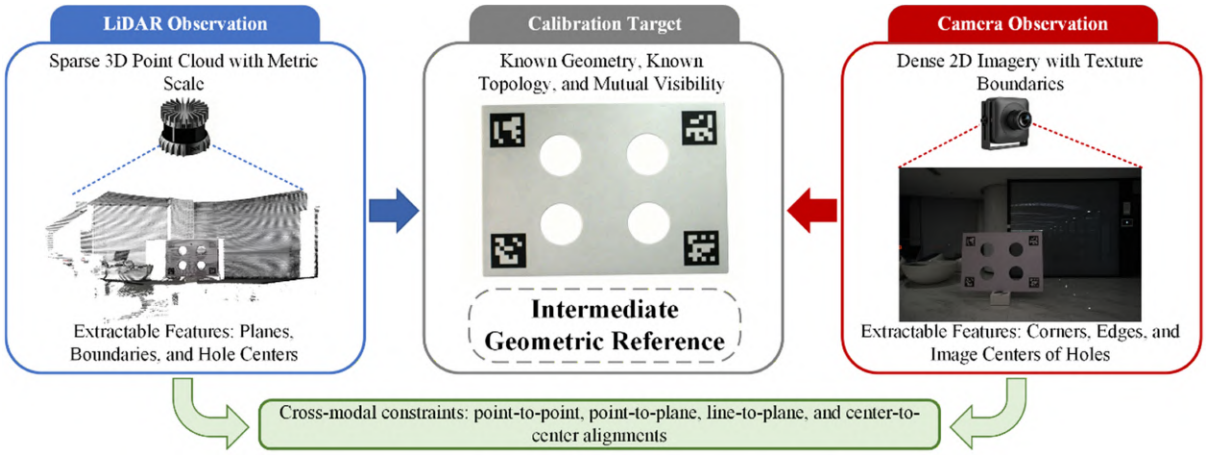

Figure 3: Artificial-target-based construction of cross-modal co-observation constraints. The calibration target serves as an intermediate geometric reference between sparse LiDAR point clouds and dense camera images.

As illustrated in Fig. 3, the role of the target is to provide an intermediate geometric reference. The same object is observed as three-dimensional structures in the LiDAR point cloud and as two-dimensional features in the camera image. The resulting constraints may take the form of point-to-point, point-to-plane, line-to-plane, center-to-center, or reprojection residuals. Early studies laid the groundwork for this paradigm. Zhang and Pless used a planar target to calibrate a camera and a laser range finder by relating laser range measurements to image observations [25]. Geiger et al. further demonstrated that checkerboard structures can support automatic calibration between cameras and range sensors from a single shot [26]. Collectively, these works showed that known artificial structures make cross-modal calibration reproducible and numerically well constrained.

3.1 Target Types and Constraint Sources

Different target designs yield different observable features and consequently lead to distinct calibration constraints. Table 2 summarizes representative target types, their feature extraction strategies, and the corresponding forms of extrinsic constraints. This comparison reveals that the difference among targets lies not merely in their physical appearance, but in how they render certain cross-modal features visible and repeatable.

Table 2: Typical artificial target types and their extrinsic constraint forms.

Target type	Features and constraints	Representative works
Planar target	Planar boards, checkerboards, and marker boards provide image corners, marker poses, plane normals, boundary lines, and line–plane constraints. Calibration is usually formulated as point-to-plane, line-to-plane, or reprojection optimization.	Zhang and Pless [25]
Planar target with holes	Circular or regular holes generate depth-discontinuity boundaries in point clouds. Hole contours and circle centers can be extracted from LiDAR data and matched with image-side circle centers or marker poses.	Beltrán et al. [27]; Yan et al. [28]
Multi-plane checkerboard	Multiple checkerboard planes provide richer pose diversity than a single plane and reduce degeneracy caused by limited viewpoint variation.	Geiger et al. [26]
Three-dimensional geometric target	Spheres, trihedrons, and polyhedral targets provide sphere centers, non-coplanar planes, intersection lines, and spatial corners, improving the observability of rotation and translation.	Gong et al. [29]; Kümmerle et al. [30]; Tóth et al. [31]
Intensity-assisted target	Reflective or high-contrast materials make the target detectable in the LiDAR intensity channel. Intensity clusters, corners, and centroids are then matched with image features.	Wang et al. [32]

From the perspective of constraint construction, planar targets primarily provide planes, edges, corners, and hole centers. Three-dimensional targets offer spatially distributed features such as sphere centers, non-coplanar planes, and intersection lines. Reflectance-intensity-assisted targets enhance LiDAR-side feature extraction by introducing material contrast. The following subsections discuss these three categories in greater detail.

3.2 Planar and Hole-Based Targets

Planar checkerboards, fiducial-marker boards, circular grids, and boards with holes are among the most common target forms. On the camera side, mature algorithms can detect sub-pixel corners, tag poses, edges, circles, or hole centers. On the LiDAR side, ordinary image texture is typically invisible, so the target must be detected through its plane, boundary, depth discontinuity, or geometric contour.

Hole-based planar targets improve the visibility of target features in sparse point clouds. Circular holes create clear depth-discontinuity boundaries, which enable the LiDAR to extract hole contours and estimate circle centers. Beltrán et al. designed a target with circular holes and formulated calibration as the registration of reference points extracted from different sensors [27]. JointCalib introduced circular holes around a checkerboard and jointly optimized camera intrinsics and LiDAR–camera extrinsics using both checkerboard and hole features [28]. FAST-Calib further emphasized fast, scan-pattern-independent extraction of hole edges and centers [33].

The workflow in Fig. 4 illustrates why hole-based targets are effective. The same circular holes are detected

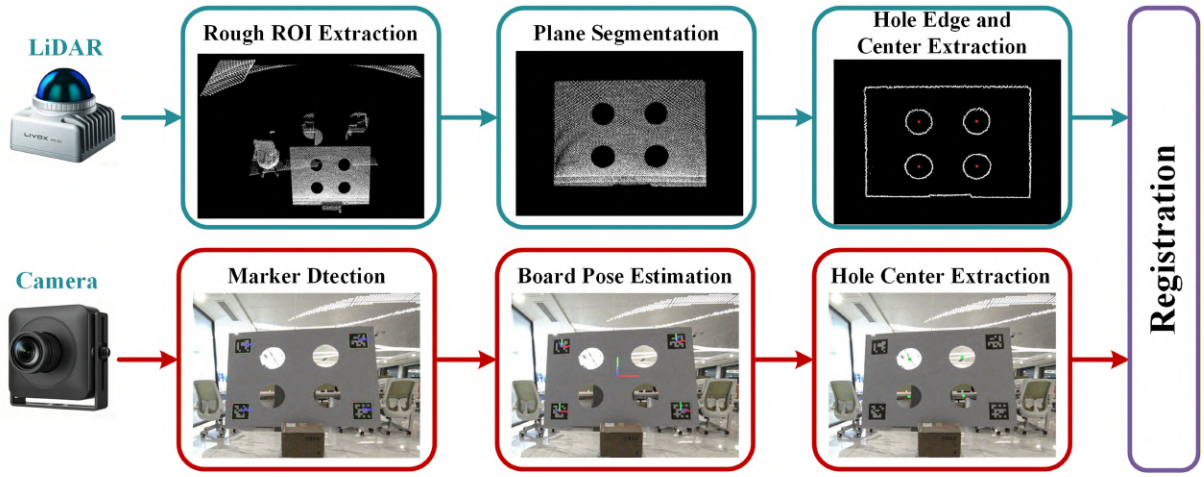

Figure 4: System pipeline of FAST-Calib. The LiDAR branch extracts the target region, segments the plane, and estimates hole centers, while the camera branch detects markers, estimates the board pose, and extracts image-side hole centers.

as depth-discontinuity structures in the point cloud and as image features in the camera. Calibration can then be formulated as the alignment of corresponding hole centers. Let $\boldsymbol{c}_{ij}^{L}$ and $\boldsymbol{c}_{ij}^{C}$ denote the j-th hole center in the i-th observation frame, expressed in the LiDAR and camera frames. The residual can be written as

$$E_{\text{hole}} = \sum_{i=1}^{M} \sum_{j=1}^{H} \left\| \boldsymbol{c}_{ij}^{C} - \left(\boldsymbol{R}_{L}^{C} \boldsymbol{c}_{ij}^{L} + \boldsymbol{t}_{L}^{C} \right) \right\|_{2}^{2}, \tag{5}$$

where M is the number of observation frames and H is the number of hole centers in each frame. This objective is a target-specific instance of the general residual minimization in Section 2.

Planar targets are attractive because they are easy to fabricate and deploy. However, a single plane provides limited constraints on certain degrees of freedom. If target poses are insufficiently diverse, or if point-cloud coverage is incomplete, the calibration may suffer from weak observability. Practical systems therefore often resort to multiple target poses, hole structures, boundary features, or multi-plane designs to strengthen the constraint distribution.

3.3 Three-Dimensional Geometric Targets

Three-dimensional targets improve observability by introducing spatial structures that are not confined to a single plane. Spheres, trihedrons, V-shaped targets, cubes, and polyhedrons provide non-coplanar planes, intersection lines, spatial corners, or view-invariant centers. Among these, spherical targets are widely used because their centers remain geometrically consistent across viewpoints.

On the LiDAR side, sphere centers are typically estimated by segmenting spherical point clusters and fitting a sphere model. On the camera side, the sphere appears as a circle or ellipse, from which the spatial center or its

projection can be inferred using the known radius and camera intrinsics. Kümmerle et al. employed spherical targets for automatic calibration of multiple cameras and depth sensors [30]. Tóth et al. extended this approach to LiDAR–camera calibration by registering sphere-center observations estimated from point clouds and images [31].

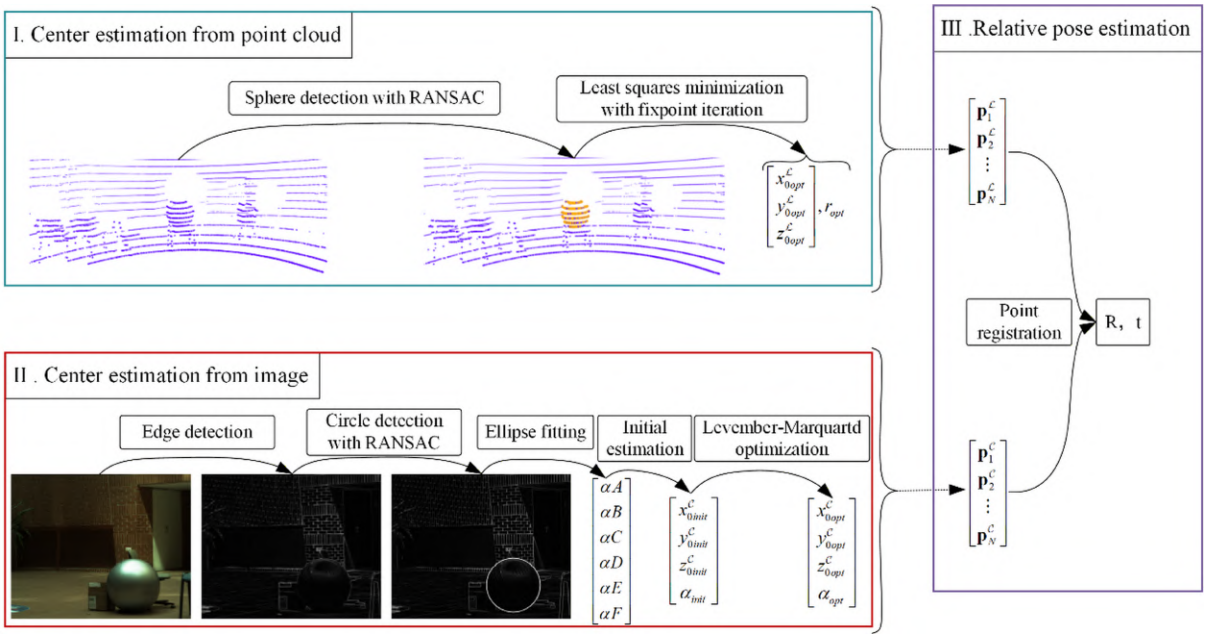

Figure 5: Pipeline of spherical-target-based extrinsic calibration. Sphere centers are estimated separately from point clouds and images, and the relative pose is solved by registering the corresponding center sets.

The pipeline in Fig. 5 separates the problem into center estimation and pose estimation. Let $\boldsymbol{q}_k^L$ denote a LiDAR point sampled from a spherical target, $\boldsymbol{o}^L$ the sphere center, and r the sphere radius. The LiDAR-side sphere fitting problem can be expressed as

$$\left(\boldsymbol{o}^L, r\right)^* = \arg\min_{\boldsymbol{o}^L, r} \sum_k \left(\left\|\boldsymbol{q}_k^L - \boldsymbol{o}^L\right\|_2 - r\right)^2. \tag{6}$$

Once sphere centers are estimated from both modalities, the extrinsic parameters can be solved by registering the corresponding center sets. Compared with planar boards, spherical targets are less sensitive to viewpoint changes; however, their accuracy depends on reliable sphere segmentation and sufficient point-cloud coverage.

Trihedral and polyhedral targets provide a complementary form of spatial constraint. They contain multiple non-coplanar planes, intersection lines, and spatial corners, which can mitigate degeneracy associated with single-plane calibration. Gong et al. formulated LiDAR–camera calibration using an arbitrary trihedron and exploited the geometric constraints of three intersecting planes [29]. These targets are more complex to manufacture and deploy than planar boards, but they can enhance stability when richer three-dimensional constraints are required.

3.4 Reflectance-Intensity-Assisted Targets

Reflectance-intensity-assisted targets exploit the LiDAR intensity channel in addition to geometric shape. By attaching reflective materials or designing patterns with strong reflectance contrast, the target becomes separable not only in camera images but also in LiDAR intensity measurements. This advantage is particularly valuable because ordinary checkerboard textures are clearly visible to cameras yet are not directly observable by LiDAR geometry alone.

Wang et al. proposed a reflectance-intensity-assisted checkerboard for automatic calibration between a 3D LiDAR and a panoramic camera [32]. Their method segments the checkerboard region from the point cloud, estimates checkerboard corners using intensity information, detects corresponding corners in the panoramic image, and then refines the extrinsic parameters through nonlinear optimization.

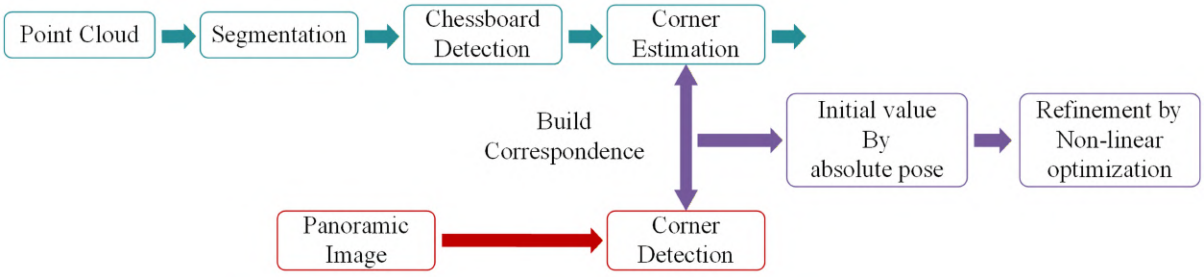

Figure 6: Pipeline of reflectance-intensity-assisted checkerboard calibration. LiDAR intensity helps extract checkerboard features from point clouds, which are then matched with image-side corners for pose estimation and nonlinear refinement.

As shown in Fig. 6, intensity is mainly used to improve feature extraction on the LiDAR side. The final calibration constraint is still geometric. Let $\boldsymbol{p}_i^L$ be the i-th three-dimensional feature extracted from LiDAR, and let $\boldsymbol{x}_i = [u_i, v_i]^\mathsf{T}$ be its corresponding image feature. The reprojection residual can be written as

$$E_{\text{repr}} = \sum_i \rho\left(\left\|\boldsymbol{x}_i - \pi\left(\boldsymbol{K}\left(\boldsymbol{R}_L^C \boldsymbol{p}_i^L + \boldsymbol{t}_L^C\right)\right)\right\|_2^2\right), \tag{7}$$

where $\pi(\cdot)$ denotes perspective projection and $\rho(\cdot)$ is a robust loss. This formulation makes the role of intensity clear: it helps obtain reliable LiDAR-side features, while the extrinsic parameters are still estimated through reprojection consistency.

Intensity-assisted targets can alleviate the difficulty of detecting checkerboard-like features in sparse point clouds. However, LiDAR intensity is affected by material reflectance, observation distance, incidence angle, and sensor-specific response. Therefore, intensity is better regarded as an auxiliary cue for feature extraction rather than a universally stable constraint source.

In summary, target-based methods provide highly accurate and reproducible solutions for LiDAR-camera ex-

trinsic calibration by introducing deliberately designed geometric or reflective features. Their primary strengths lie in procedural controllability and the physical interpretability of the resulting constraints [34]. However, the applicability of this paradigm is inherently constrained by the requirement for manual deployment of calibration targets, which renders it inadequate for online calibration needs arising from long-term sensor operation under vibration, thermal deformation, or mechanical displacement. This limitation has progressively shifted research focus from reliance on artificial targets toward targetless approaches that leverage intrinsic structures in natural scenes, which will be elaborated in the following chapter.

4 Targetless Methods: Explicit Cross-Modal Matching in Natural Scenes

Targetless LiDAR–camera calibration removes the requirement for manually deployed calibration objects and instead constructs constraints from naturally occurring scene evidence. Compared with target-based methods, the main challenge is that natural scenes do not provide predefined geometry, known dimensions, or fixed topological structures. Therefore, targetless methods must first identify visual and geometric evidence observable by both sensors, and then convert this evidence into explicit constraints on the extrinsic parameters.

This section organizes targetless explicit matching methods according to the space in which effective correspondences are constructed. As summarized in Table 3, three categories are commonly observed. In 2D–3D methods, image features are matched with LiDAR-side points, lines, planes, objects, or semantic structures before projection. In 3D–3D methods, camera-side depth, reconstructed geometry, or motion is lifted to 3D and then registered with LiDAR observations. In 2D–2D methods, LiDAR measurements are first projected or rendered into the image plane, and calibration is performed by aligning two image-domain representations. Learning modules and large vision models may appear in these methods, but they are treated here as feature, mask, or correspondence extractors when the final extrinsic estimation is still solved by explicit geometric or statistical optimization.

This section focuses on methods that ultimately solve for extrinsic parameters through explicit geometric or statistical optimization. Even when deep learning modules are employed at the front end for feature extraction, semantic segmentation, or correspondence prediction, a method remains within the scope of this chapter as long as the final pose estimation is accomplished by explicit optimization procedures such as PnP solvers, ICP registration, edge alignment optimization, or mutual information maximization. Taking CMRNext, CFNet, and DXQ-Net as representative examples, these approaches predict calibration flows or matching weights through networks, yet their final pose estimation still relies on EPnP combined with RANSAC or similar geometric solvers, and are therefore

systematically discussed in this section. In contrast, Chapter 5 is dedicated to end-to-end regression strategies and implicit feature alignment methods that do not depend on such explicit solvers.

Table 3: Targetless explicit matching methods grouped by correspondence space.

Correspondence space	Typical evidence	Main calibration objective
2D–3D	Point–pixel matches, image lines, LiDAR planes, semantic objects, and semantic edges	PnP, reprojection minimization, point-to-line residuals, point-to-plane residuals, and semantic consistency
3D–3D	Stereo or SfM point clouds, semantic 3D landmarks, reconstructed meshes, and sensor ego-motion	Rigid registration, ICP-style alignment, semantic landmark registration, and hand–eye calibration
2D–2D	Projected LiDAR depth, reflectance, edges, masks, semantic maps, and statistical image representations	Edge alignment, boundary-mask consistency, mutual information, normalized information distance, and semantic distribution alignment

4.1 2D–3D Correspondences

2D–3D methods preserve the LiDAR-side primitive as a three-dimensional entity when the correspondence is formed. The camera model then connects this 3D primitive to image measurements. A typical point-level residual can be written as

$$E_{\mathrm{2D3D}} = \sum_i \rho\left(\left\|\boldsymbol{x}_i - \pi\left(\boldsymbol{K}\left(\boldsymbol{R}_L^C \boldsymbol{p}_i^L + \boldsymbol{t}_L^C\right)\right)\right\|_2^2\right), \tag{8}$$

where $\boldsymbol{p}_i^L$ is a LiDAR-side 3D feature, $\boldsymbol{x}_i$ is the corresponding image feature, and $\rho(\cdot)$ is a robust loss. This residual is a natural-scene instance of the general optimization model introduced in Section 2.

Point and point–pixel methods directly associate image pixels with LiDAR-side 3D points. Scaramuzza et al. estimated camera–laser extrinsics from natural scenes by relating image observations to laser range measurements [7]. Recent methods retain this geometric structure while learning more reliable correspondences. CMRNext predicts point–pixel matches and then estimates pose with a PnP solver [35]. CFNet converts calibration flow into 2D–3D correspondences and applies EPnP with RANSAC [14]. DXQ-Net further introduces quality-aware flow and differentiable pose estimation [15]. These methods are compact and interpretable, but their performance depends on overlap, correspondence confidence, and outlier rejection.

Structural 2D–3D methods use lines, edges, planes, or directions as correspondence carriers. Representative matching visualizations for point correspondences, line/edge matching and plane constraints are shown in Fig. 7. Zhou et al. employed natural line and plane correspondences for camera–LiDAR calibration [8]. VP-Calib estimates rotation from image vanishing points and 3D dominant line directions [36]. LCC extracts LiDAR-side edge or line structures through voxel cutting and plane fitting [9], while MLCC extends this idea to multiple small-FoV LiDARs and cameras using adaptive voxelization [10]. The complete MLCC workflow is illustrated in Fig. 8.

Other structural methods further enrich this category. Galibr uses ground-plane initialization followed by edge-based refinement [37]. PLK-Calib represents LiDAR lines with Plücker coordinates and constructs line constraints

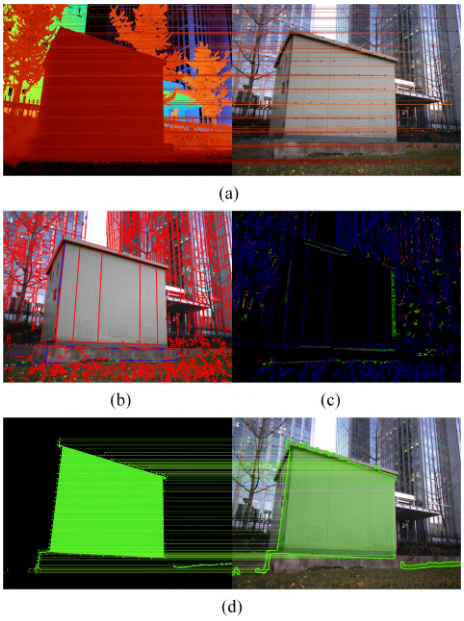

Figure 7: Representative 2D–3D feature matching processes in targetless calibration, including point correspondences, line or edge matching, and plane-based constraints.

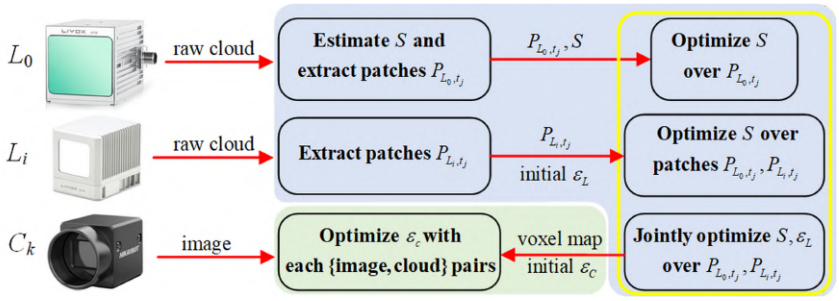

Figure 8: Overall workflow of MLCC. Local patches are extracted from LiDAR point clouds and camera images, followed by extrinsic refinement over the selected patch correspondences.

for single-shot calibration [38]. YOCO employs plane orientation and coplanar constraints as primary geometric evidence [39]. MFCalib combines depth-continuous, depth-discontinuous, and intensity-discontinuous LiDAR edges for single-shot calibration [40]. Earlier line-based and geometry-based studies also explored natural line correspondences, auxiliary-object-free calibration, and line-based automatic calibration [41–43]. PBACalib formulates targetless calibration as plane-constrained bundle adjustment for high-resolution LiDAR–camera systems [44].

For line and plane primitives, the residual often changes from point reprojection to structural distance. For example, if ℓ_i is an image line and $\boldsymbol{p}_{ij}^L$ is a LiDAR point associated with that line after projection, a point-to-line residual can be written as

$$E_{\text{line}} = \sum_{i,j} \rho \left(d \left(\ell_i, \pi \left(\boldsymbol{K} \left(\boldsymbol{R}_L^C \boldsymbol{p}_{ij}^L + \boldsymbol{t}_L^C \right) \right) \right)^2 \right), \tag{9}$$

where $d(\cdot, \cdot)$ denotes the distance from a projected point to an image line. Such methods are effective in urban

roads, building facades, corridors, and other structured scenes. However, repeated parallel facades, dominant ground planes, short visible line segments, or weak field-of-view overlap can make some rotation or translation components poorly constrained.

Semantic 2D–3D methods use objects, semantic centroids, or semantic edges as correspondence carriers. SOIC converts initialization into a PnP problem using semantic centroids, followed by semantic consistency refinement [45]. ATOP employs attention-based object matching to generate 2D–3D object correspondences before optimization [46]. SE-Calib extracts semantic edges from point clouds and images and maximizes semantic response consistency [47]. These methods reduce reliance on hand-crafted low-level edges, but their accuracy depends on semantic detection quality, static objects, and sufficient common field of view.

4.2 3D–3D Correspondences

3D–3D methods first construct camera-side 3D information or camera-side motion estimates, and then solve a 3D registration or hand–eye calibration problem. Compared with 2D–3D methods, the main advantage is that calibration can be expressed in a common three-dimensional space. The main limitation is that camera-side 3D reconstruction or odometry must be sufficiently accurate and metric.

One line of work reconstructs camera-side geometry before registration. Dhall et al. formulated LiDAR–camera calibration using 3D–3D point correspondences [48]. Wang et al. estimated extrinsics between a monocular camera and LiDAR in natural scenes using recovered 3D structure [49]. Li et al. registered panoramic image sequences and mobile laser scanning data using semantic 3D features [50]. Nagy and Benedek generated camera-side point clouds with structure from motion and registered them with LiDAR point clouds for on-the-fly calibration [51]. TEScalib employed stereo-camera mesh reconstruction and point-cloud registration for targetless LiDAR–stereo calibration [52]. Song and Lee further proposed a voxel-based network for online self-calibration of 3D measurement sensors [53].

When corresponding 3D structures are available, the residual can be written as

$$E_{\mathrm{3D3D}} = \sum_i \left\| \boldsymbol{q}_i^C - \left(\boldsymbol{R}_L^C \boldsymbol{p}_i^L + \boldsymbol{t}_L^C \right) \right\|_2^2, \tag{10}$$

where $\boldsymbol{q}_i^C$ is a camera-side 3D point, landmark, or reconstructed feature corresponding to the LiDAR feature $\boldsymbol{p}_i^L$. This formulation benefits from mature 3D registration tools, but it is sensitive to reconstruction scale, textureless regions, sparse stereo depth, and local minima in registration.

A second line of 3D–3D methods uses relative sensor motion as the correspondence. This leads to a hand–eye calibration formulation. Let $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ denote relative motions estimated from the camera and LiDAR trajectories,

and let $\boldsymbol{X}$ denote the unknown rigid transform between the two sensor frames. The motion-consistency constraint is

$$\boldsymbol{A}_k \boldsymbol{X} = \boldsymbol{X} \boldsymbol{B}_k. \tag{11}$$

Dual-quaternion formulations provide a classical solution to this problem [54]. Ishikawa et al. estimated LiDAR–camera calibration from motions obtained by sensor-fusion odometry [55]. Park et al. further proposed a targetless and structureless spatiotemporal calibration method based on motion consistency [56]. Motion-based methods are useful when individual frames contain weak structure; however, they require sufficiently excited motion, accurate odometry, and careful handling of temporal offset, rolling shutter, and LiDAR motion distortion.

4.3 2D–2D Correspondences

2D–2D methods first transform LiDAR measurements into an image-plane representation and then compare this representation with camera-side evidence. This formulation places both modalities in a common coordinate system and is therefore suitable for edge maps, boundary masks, semantic masks, reflectance images, depth maps, and statistical similarity measures. A general projection process can be expressed as

$$\mathcal{M}_L = \mathcal{S}\left(\left\{\pi\left(\boldsymbol{K}\left(\boldsymbol{R}_L^C \boldsymbol{p}_i^L + \boldsymbol{t}_L^C\right)\right), a_i\right\}_i\right), \tag{12}$$

where a_i denotes a LiDAR attribute such as depth, reflectance, semantic label, or normal direction, and $\mathcal{S}(\cdot)$ denotes z-buffering, accumulation, interpolation, or kernel splatting.

Point, pixel, edge, and boundary methods compare image-plane evidence after LiDAR measurements have been projected or accumulated into the camera view. SPTL-LCC constructs dense pixel-level correspondences in this shared image plane and then estimates extrinsics through explicit optimization [57]. Castorena et al. formulated camera–LiDAR autocalibration as image-plane edge alignment [58], and later introduced a motion-guided extension with accelerated depth upsampling [59]. Zhu et al. employed LiDAR reflectivity to construct targetless edge-to-edge constraints [60]. Levinson and Thrun performed automatic online calibration by aligning camera and laser observations over time [61]. Other edge or boundary alignment methods leverage Gaussian mixture models, coarse-to-fine edge matching, motion initialization, or class-agnostic boundary masks [62–65].

For image-plane boundary alignment, a typical residual can be written as

$$E_{\text{bdry}} = \sum_{\boldsymbol{u} \in \mathcal{B}_L} D_C(\boldsymbol{u})^2, \tag{13}$$

where $\mathcal{B}_L$ is the projected LiDAR boundary set and $D_C(\boldsymbol{u})$ is the distance transform of the camera-side edge or boundary map. These objectives can be accurate when strong contours exist, but shadows, texture edges, dynamic

objects, and sparse projections may introduce misleading evidence.

Mutual-information and statistical image methods avoid explicit feature matching. Instead, they optimize a similarity measure between a camera image and a projected LiDAR attribute image. Taylor and Nieto employed normalized mutual information for automatic LiDAR–camera calibration. Taylor et al. introduced a gradient-orientation measure for multimodal calibration [66]. Irie et al. used a bagged dependence estimator for targetless calibration [67]. Napier et al. studied cross-calibration of push-broom LiDARs and cameras in natural scenes [68]. Ishikawa et al. optimized an intensity-variance cost [69], while Pandey et al. maximized mutual information between camera intensity and LiDAR reflectance [11]. Guislain et al. registered urban LiDAR point sets to images using LiDAR-derived reflectance and surface cues [70]. Koide et al. proposed a general single-shot targetless calibration toolbox using image-plane matching and normalized information distance [71]. SemCal further introduced semantic information and a neural mutual-information estimator [72].

A common normalized mutual-information objective is

$$\left(\boldsymbol{R}_L^C, \boldsymbol{t}_L^C\right)^* = \arg\max_{\boldsymbol{R}_L^C, \boldsymbol{t}_L^C} \frac{H(\mathcal{I}_C) + H(\mathcal{I}_L)}{H(\mathcal{I}_C, \mathcal{I}_L)}, \tag{14}$$

where $\mathcal{I}_C$ is the camera image or feature image, $\mathcal{I}_L$ is the projected LiDAR attribute image, and $H(\cdot)$ denotes entropy. These objectives are flexible but often non-convex. Their performance depends on initialization, overlap, reflectance stability, illumination, histogram construction, and the discriminability of the shared image-domain evidence.

Semantic and mask-based 2D–2D methods compare projected LiDAR semantic or mask evidence with image-side segmentation. SST-Calib employs semantic masks on both modalities and optimizes a pixel-wise bidirectional loss for spatial and temporal calibration [73]. Zhu et al. employed semantic information for online camera–LiDAR calibration [74]. Calib-Anything uses Segment Anything to generate image masks and maximizes the consistency of projected LiDAR points within these masks [75]. Its workflow is illustrated in Fig. 9.

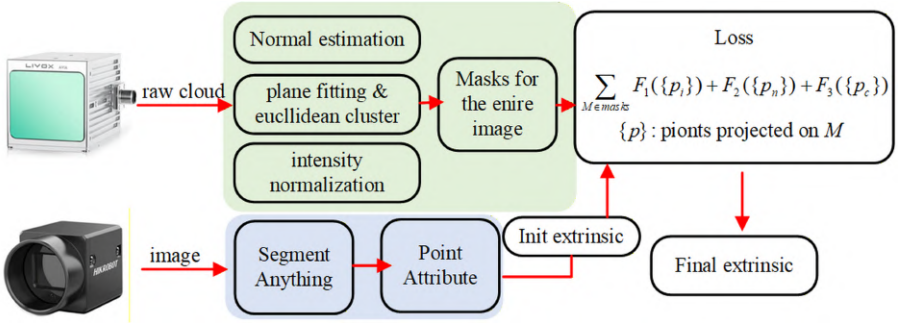

Figure 9: Overall workflow of Calib-Anything. Segment Anything provides image masks, while LiDAR-derived attributes are projected into the image domain to optimize mask-level consistency.

Recent methods further exploit large vision models and semantic distribution alignment. MIAS-LCEC leverages

large vision models to generate cross-modal mask matches for online target-free calibration [76]. Semantic distribution alignment constructs camera and LiDAR semantic probability fields in the image domain and optimizes distributional consistency on $SE(3)$ [77]. These methods are attractive in cluttered scenes because masks and semantic regions are more stable than thin edges. However, segmentation failure, category mismatch, partial occlusion, moving objects, and prompt sensitivity can directly affect the calibration objective.

In summary, targetless explicit matching methods replace artificial calibration targets with inherent geometric, semantic, or statistical evidence from natural scenes, substantially enhancing the deployment convenience of the calibration pipeline. However, the intrinsic limitations of these approaches should not be overlooked. 2D-3D methods are highly sensitive to feature sparsity and scene degeneracy. 3D-3D methods rely critically on the accuracy of reconstruction or odometry estimation. 2D-2D methods are hampered by projection sparsity, non-convex optimization objectives, and semantic segmentation errors. Fundamentally, the methods above still depend on manually predefined cue types such as edges, planes, and mutual information as constraint sources, whose effectiveness becomes precarious under feature-deficient or highly variable scene conditions. In this context, a natural evolutionary path is to automatically learn cross-modal correspondences through data-driven paradigms, thereby eliminating dependence on specific handcrafted priors. This constitutes the core motivation of the learning-based paradigm elaborated in Section 5. Furthermore, Section 6 will demonstrate how differentiable rendering elevates the calibration problem from local correspondence consistency to global scene consistency, marking yet another paradigm shift in this field.

5 Learning-Based Methods: Cross-Modal Alignment and Online Calibration

The preceding sections reviewed target-based calibration and targetless explicit matching in natural scenes. These methods construct constraints from manually designed targets or explicitly selected scene cues, such as corners, planes, edges, lines, semantic regions, or statistical image similarities. Although such constraints are interpretable, their effectiveness depends strongly on feature visibility, scene structure, sensor overlap, and initialization quality. In open environments, particularly in autonomous driving and mobile robotics, these assumptions are often difficult to satisfy.

It should be clarified that Chapter 4 has systematically reviewed hybrid methods that combine deep learning-based feature extraction with classical explicit geometric solvers, including PnP, ICP, and edge alignment optimization. Approaches such as CFNet, DXQ-Net, and CMRNext, though utilizing networks to predict correspondences or

calibration flows, ultimately rely on EPnP with RANSAC or other external solvers for final pose estimation, and are therefore not revisited in this chapter.

In contrast, the present chapter concentrates on end-to-end learning paradigms that operate without dependence on such external explicit solvers.Learning-based calibration methods address these limitations by replacing manually designed correspondences with learned cross-modal representations. Instead of explicitly defining which point, edge, or object should be matched, a network learns feature correspondences, calibration flow, geometric consistency, or extrinsic corrections from image and point-cloud data. This shifts the calibration problem from rule-based feature construction to data-driven cross-modal alignment. The primary motivation is online calibration: after deployment, sensor extrinsics may drift due to vibration, temperature variation, mechanical deformation, or reinstallation. A practical calibration method should therefore detect and correct extrinsic errors without relying on dedicated targets.

Existing learning-based LiDAR–camera calibration methods can be grouped into four categories, as summarized in Table 4: direct parameter regression, cross-modal feature matching, geometric-constraint fusion, and iterative or multi-sensor refinement. These categories are not strictly independent. Many recent methods combine learned features with explicit geometric solvers, which forms a bridge between the targetless methods in Section 4 and the differentiable scene-representation methods in Section 6.

Table 4: Main categories of learning-based LiDAR–camera calibration methods.

Category	**Core idea**	**Representative methods**
Direct parameter regression	Regress the 6-DoF extrinsic correction directly from RGB images and projected LiDAR maps.	RegNet, CalibNet, RGGNet, FusionNet, NetCalib [12, 16, 78–80]
Cross-modal feature matching	Learn image–LiDAR correspondences, cost volumes, or calibration flow before solving extrinsics.	CMRNet, CMRNext, LCCNet, CFNet, DXQ-Net, TransCalib, EdgeCalib [13–15, 35, 81, 82]
Geometric-constraint fusion	Embed reprojection, depth, semantic, or consistency losses into network training.	CalibRCNN, RobustCalib, CalibFormer, RLCFormer [83–86]
Iterative and multi-sensor calibration	Refine extrinsics in multiple stages or extend learning-based calibration to variable sensor suites.	DualTrans-Calib, SPTL-LCC, Simple-Calib, Multi-Calib, LiREC-Net [57, 87–90]

5.1 Direct Parameter Regression

Direct regression is the earliest and simplest learning-based paradigm. It takes an RGB image and a projected LiDAR representation (e.g., a depth map or front-view map) as input and directly outputs the extrinsic correction. The network thus treats calibration as a supervised or self-supervised regression problem on $SE(3)$.

Let $\mathcal{F}_\theta(\cdot)$ denote a neural network, $\mathcal{I}$ the RGB image, and $\mathcal{D}_L$ the projected LiDAR depth map. A direct regression model can be written as

$$\Delta\boldsymbol{\xi} = \mathcal{F}_\theta(\mathcal{I}, \mathcal{D}_L), \tag{15}$$

where $\Delta\boldsymbol{\xi}$ denotes the predicted extrinsic correction in a six-dimensional pose parameterization. A typical

supervised loss combines translation and rotation errors:

$$\mathcal{L}_{\text{reg}} = \|\Delta \boldsymbol{t} - \Delta \boldsymbol{t}^*\|_1 + \lambda \, \|\Delta \boldsymbol{r} - \Delta \boldsymbol{r}^*\|_1 \,, \tag{16}$$

where $\Delta \boldsymbol{t}$ and $\Delta \boldsymbol{r}$ are the predicted translation and rotation corrections, and λ balances their scales.

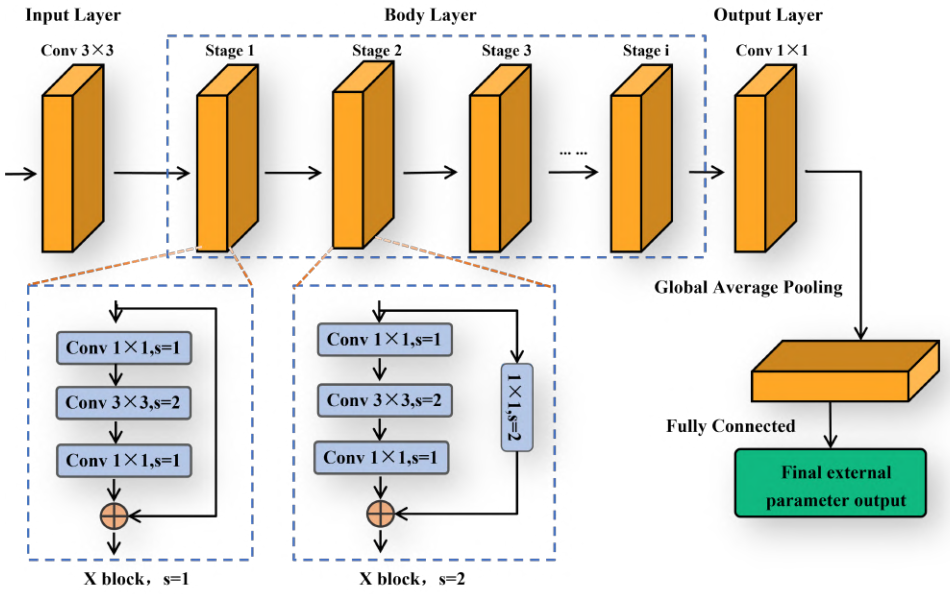

Figure 10: Architecture of a direct regression network for LiDAR–camera extrinsic calibration. RGB and LiDAR-derived representations are encoded and fused before the final extrinsic parameter output.

RegNet is a representative early instance of this paradigm [78]. CalibNet introduced geometric supervision based on 3D spatial transformation and depth consistency [12]. RGGNet further enforced rotation prediction on the $SO(3)$ manifold to improve robustness [16]. FusionNet and NetCalib enhanced feature extraction through hierarchical fusion and multi-scale representations [79, 80]. As illustrated in Fig. 10, the advantage of direct regression lies in its straightforward inference pipeline and rapid deployment. However, because the network outputs a black-box pose correction, these methods often lack explicit geometric interpretability and can generalize poorly when the scene, sensor configuration, or initial miscalibration range deviates from the training data.

5.2 Cross-Modal Feature Matching

Cross-modal feature matching methods retain the geometric spirit of classical calibration while replacing hand-crafted correspondences with learned ones. Instead of directly regressing extrinsics, the network predicts a matching cost, correspondence field, or calibration flow between image features and projected LiDAR features. The final pose is then estimated from the learned correspondences or refined by a differentiable geometric module.

A representative formulation is cost-volume construction. Let $\boldsymbol{f}_I(\boldsymbol{u})$ be the image feature at pixel $\boldsymbol{u}$, and let

$\boldsymbol{f}_L(\boldsymbol{u}+\boldsymbol{\delta})$ be the LiDAR-derived feature at an offset location. The matching cost can be expressed as

$$C(\boldsymbol{u},\boldsymbol{\delta}) = \frac{\boldsymbol{f}_I(\boldsymbol{u})^{\mathsf{T}} \boldsymbol{f}_L(\boldsymbol{u}+\boldsymbol{\delta})}{\|\boldsymbol{f}_I(\boldsymbol{u})\|_2 \, \|\boldsymbol{f}_L(\boldsymbol{u}+\boldsymbol{\delta})\|_2}. \tag{17}$$

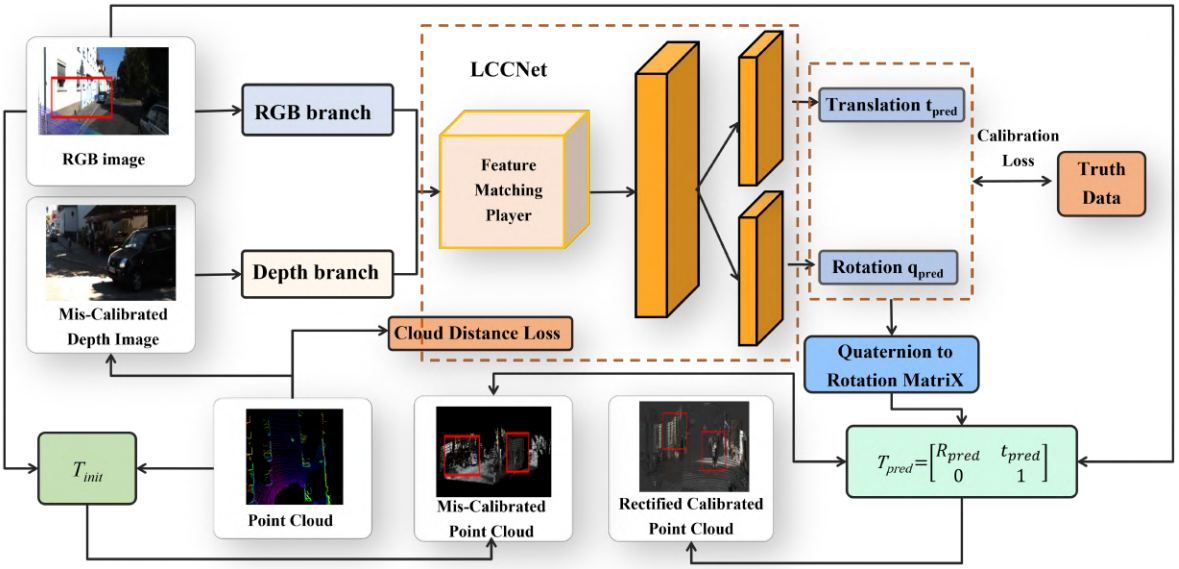

Figure 11: LCCNet framework. RGB and LiDAR-derived depth features are matched through a cost-volume-like module, and the predicted rotation and translation corrections are supervised by calibration and cloud-distance losses.

Currently, cross-modal feature alignment methods can be categorized into three mainstream technical routes. The first constructs differentiable matching cost volumes within the network, exemplified by LCCNet, which explicitly models pixel-level spatial correlations by building a cost volume between RGB and projected depth features and directly regresses pose corrections from the cost field, enabling end-to-end calibration without external solvers [13]. The framework of LCCNet is illustrated in Figure 11.The second embeds differentiable 3D spatial transformation layers into the network, represented by CalibNet, which projects LiDAR points onto the image plane according to the current pose and computes geometric losses, with gradients back-propagated to update network weights, thereby equipping the network with geometric reasoning capability during training [12]; building upon this work, RGGNet introduces SO(3) manifold constraints to enforce orthogonality of rotation predictions, significantly enhancing numerical robustness under noisy conditions [16]. The third leverages Transformer attention mechanisms for implicit cross-modal feature alignment, typified by the unsupervised dual-Transformer alignment method, which aligns image and LiDAR features via dual-Transformer architectures and directly regresses the extrinsic transformation without explicit correspondences [87]; RLCFormer further adapts this framework for roadside large-field-of-view scenarios through improved positional encoding and sparse attention mechanisms, enhancing calibration accuracy for distant points while preserving end-to-end deployment convenience [85].

Compared with direct regression, feature-matching methods provide better geometric interpretability because

pose estimation is linked to local or pixel-level correspondences. Their main limitation is that learned correspondences can still fail under occlusion, repeated structures, dynamic objects, or insufficient image–LiDAR overlap.

The three technical routes described above all operate without external geometric solvers, fundamentally distinguishing them from the hybrid methods in Section 4.1, where final pose estimation still relies on explicit solvers such as PnP or ICP. The advantages of these approaches lie in their complete gradient flow and efficient deployment, while circumventing the risk of numerical instability that explicit solvers may incur under anomalous inputs. The trade-off, however, is reduced physical interpretability and stronger dependence of pose estimation accuracy on training data, with robustness potentially inferior to the hybrid schemes in Section 4 under extreme viewpoints or unseen sensor configurations.

5.3 Geometric-Constraint Fusion

Geometric-constraint fusion methods aim to enhance the physical plausibility of learning-based calibration. Instead of relying solely on regression loss, they incorporate reprojection consistency, depth consistency, semantic consistency, or appearance consistency into the training objective. This design reduces the network's dependence on a single pose label and offers more robust supervision in weakly textured or partially occluded scenes.

A typical loss combines pose supervision with geometry-aware terms:

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + \alpha \mathcal{L}_{\text{repr}} + \beta \mathcal{L}_{\text{depth}} + \gamma \mathcal{L}_{\text{cons}}, \tag{18}$$

where $\mathcal{L}_{\text{repr}}$ measures reprojection consistency, $\mathcal{L}_{\text{depth}}$ measures depth consistency, and $\mathcal{L}_{\text{cons}}$ denotes appearance, semantic, or feature consistency.Figure 12 presents a typical overall framework for geometric-constraint fusion methods, which combine learned representations with multiple geometric consistency losses.

CalibRCNN employs recurrent convolutional prediction and geometric constraints for LiDAR–camera calibration [83]. RobustCalib enhances robustness through appearance and geometric consistency learning [84]. CalibFormer leverages transformer-based feature aggregation and correlation modeling for automatic calibration [85]. RLCFormer extends transformer-based calibration to roadside LiDAR–camera systems [86]. These methods demonstrate that geometry-aware supervision can improve generalization compared with pure regression. However, their performance still relies on the quality of pseudo-depth, semantic labels, projected LiDAR density, and representativeness of training data.

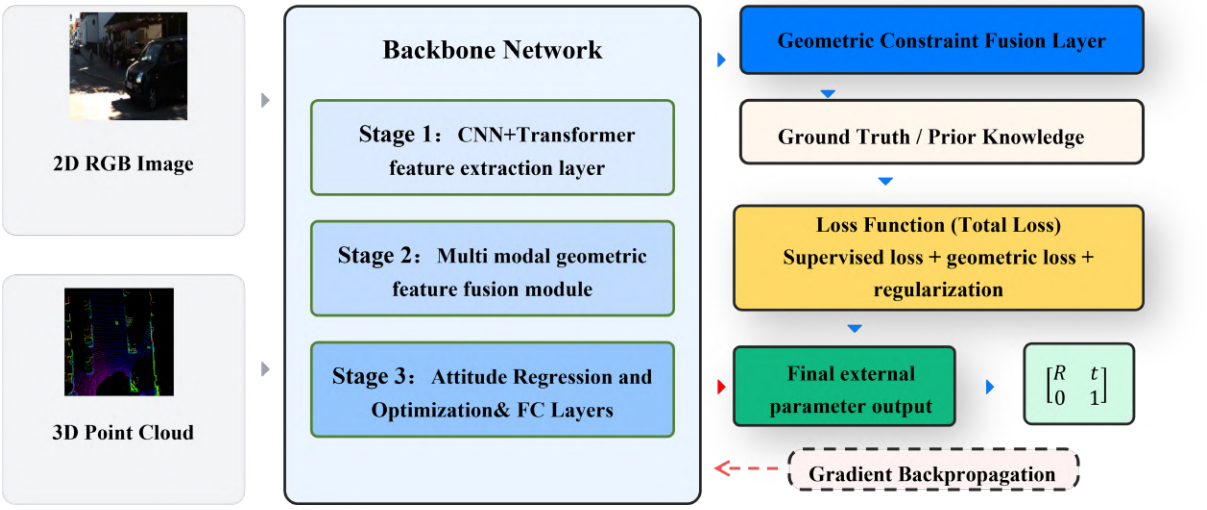

Figure 12: Overall framework of geometric-constraint fusion methods. Learned representations are constrained by geometric consistency, supervised losses, and regularization before producing the final extrinsic parameters.

5.4 Multi-Stage Refinement and Multi-Sensor Calibration

Single-step regression or matching is often insufficient when the initial extrinsic error is large. Multi-stage methods address this issue by decomposing calibration into coarse initialization and iterative refinement. At each iteration, the current extrinsic estimate is used to update the projection, recompute features or residuals, and predict a smaller correction:

$$\boldsymbol{T}_{k+1} = \Delta\boldsymbol{T}_k \boldsymbol{T}_k. \tag{19}$$

This strategy increases the capture range of learning-based calibration and reduces the risk of converging to a poor local solution.

CFNet and DXQ-Net are representative methods that combine learned offsets with geometric pose estimation [14, 15]. The architecture of CFNet is shown in Figure 13, which predicts calibration flow from RGB, depth, and point-cloud inputs, followed by geometric pose recovery and iterative refinement.CalibRefine further introduces iterative, attention-driven post-refinement for targetless online calibration. SPTL-LCC uses single-shot pixel-level correspondences within a target-free and LiDAR-type-agnostic framework [57]. Recent work has also moved beyond single LiDAR–camera pairs. Simple-Calib and Multi-Calib investigate simultaneous or scalable calibration for multiple LiDAR–camera configurations [88, 89]. LiREC-Net extends learning-based calibration to LiDAR, RGB, and event cameras [90].

In summary, learning-based methods have achieved notable progress in automation, online adaptability, and robustness to feature sparsity. However, their inherent limitations remain equally prominent. Fully supervised learning relies on expensive ground-truth annotations. Self-supervised learning is susceptible to ambiguity under

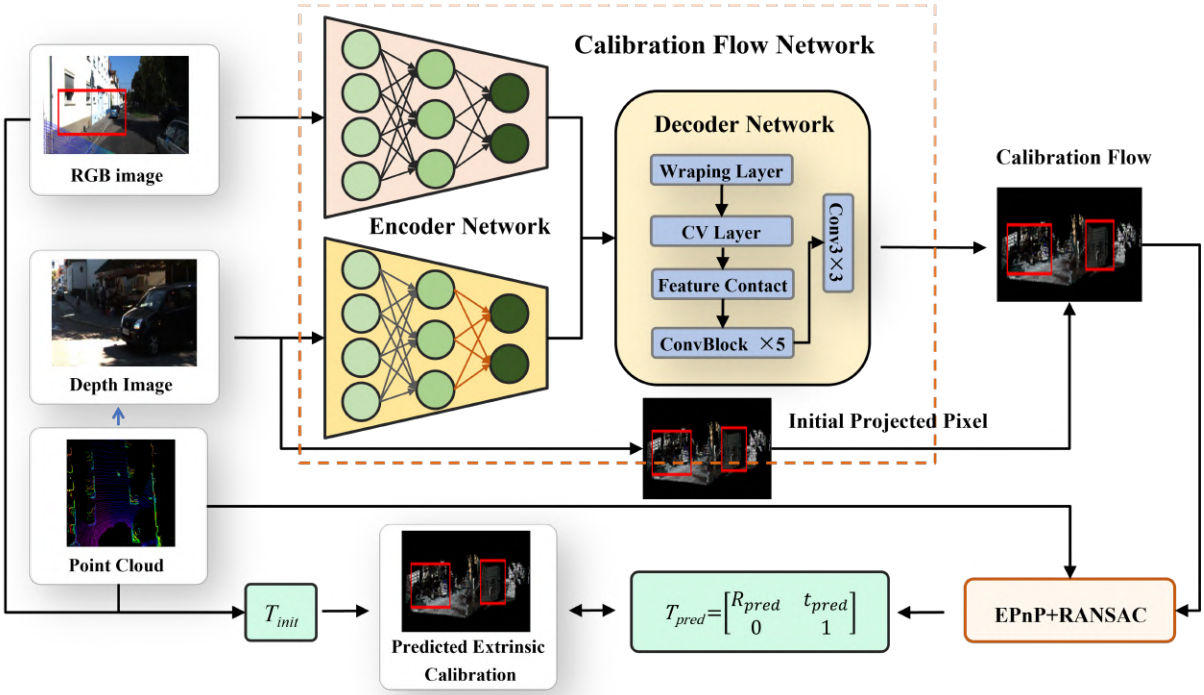

Figure 13: CFNet framework. Calibration flow is predicted from RGB, depth, and point-cloud inputs, and the final extrinsic parameters are recovered through geometric pose estimation and refinement.

occlusion, dynamic scenes, and weak overlap. Cross-domain generalization deficiency persists as an unresolved challenge. These limitations collectively drive the emergence of differentiable scene representation paradigms. This new paradigm no longer depends on local correspondences, regardless of whether they are derived from manual design or learned representations. Instead, it formulates calibration constraints within a framework of global scene rendering consistency, thereby accomplishing a fundamental transition from local constraints to global optimization.

6 Differentiable Scene Representation-Based Joint Calibration

Although the methods described in Chapters 3 to 5 differ in how correspondences are obtained, they share a common reliance on local evidence such as points, lines, planes, edges, objects, masks, or learned matching fields. In long-term open environments, factors including sparse texture, dynamic objects, occlusion, illumination variations, and limited field-of-view overlap frequently undermine the stability of such evidence. Although the methods in this chapter fall into the same targetless and data-driven categories as those in Chapters 4 and 5, the source of constraints differs fundamentally. Chapters 4 and 5 adhere to the core logic of establishing locally sparse correspondences prior to pose estimation, whereas differentiable rendering methods construct constraints through global rendering consistency, elevating the calibration problem from local feature alignment to unified

scene interpretation. However, this advancement introduces a coupling issue: the scene representation may absorb pose errors to compensate for rendering losses and converge to incorrect solutions, a problem that rarely occurs in the methods of Chapter 4 where correspondences and pose estimation are decoupled. Moreover, the computational cost substantially exceeds that of the methods in Chapters 4 and 5. How to suppress coupling and reduce overhead while leveraging the advantages of global consistency remains a central challenge for the practical deployment of this paradigm.

Differentiable scene representation-based methods shift the calibration problem from local correspondence estimation to unified scene explanation. The central premise is that camera images and LiDAR point clouds are observations of the same physical scene. If the extrinsic parameters are correct, the two modalities should be jointly explainable by a shared 3D representation through differentiable rendering. If the extrinsics are inaccurate, the inconsistency manifests as photometric residuals, depth errors, visibility conflicts, or geometric misalignment. Calibration can therefore be solved by jointly optimizing the scene representation and the sensor parameters.

A general objective can be written as

$$\boldsymbol{T}^* = \arg\min_{\boldsymbol{T},\mathcal{S}} \mathcal{L}_{\text{photo}}\left(\mathcal{R}_I(\mathcal{S},\boldsymbol{T}),\mathcal{I}\right) + \lambda_d \mathcal{L}_{\text{depth}}\left(\mathcal{R}_D(\mathcal{S},\boldsymbol{T}),\mathcal{P}\right) + \lambda_g \mathcal{L}_{\text{geom}} + \lambda_r \mathcal{L}_{\text{reg}}, \tag{20}$$

where $\mathcal{S}$ denotes the scene representation, $\mathcal{R}_I$ and $\mathcal{R}_D$ denote differentiable image and depth rendering, $\mathcal{I}$ is the camera image, and $\mathcal{P}$ is the LiDAR observation. This formulation does not assume a specific representation. $\mathcal{S}$ may be an implicit radiance field, a 3D Gaussian field, or a 2D Gaussian surface representation.

6.1 Implicit Radiance-Field Calibration

Implicit radiance fields, exemplified by NeRF, model a scene as a continuous density and color field and synthesize images through ray sampling and volume rendering [17]. BARF further demonstrated that radiance fields can be coupled with bundle-adjustment-like pose optimization, employing coarse-to-fine positional encoding to mitigate local optima caused by inaccurate poses [18]. Although BARF was originally developed for camera pose optimization, it provides a methodological foundation for treating extrinsic calibration as a differentiable rendering problem.

In LiDAR–camera calibration, implicit radiance fields transform heterogeneous observations into constraints within a shared continuous volume. LiDAR supplies sparse but metric geometry, while cameras provide dense color and texture. Instead of matching a limited set of handcrafted features, the optimization asks whether the same scene field can simultaneously explain LiDAR-supported geometry and image appearance.

INF is an early instance of this paradigm. It constructs an implicit field in which LiDAR contributes geometric

density and camera images provide photometric constraints [21]. MOISST extends implicit-scene optimization to multi-LiDAR and multi-camera systems by jointly estimating spatial extrinsics and temporal offsets [22]. SOAC improves robustness by emphasizing spatiotemporal overlap and visibility-aware constraints [23]. UniCal further introduces a drive-and-calibrate framework for fleet-scale multi-camera and multi-LiDAR calibration using differentiable volume rendering and surface alignment loss [91].

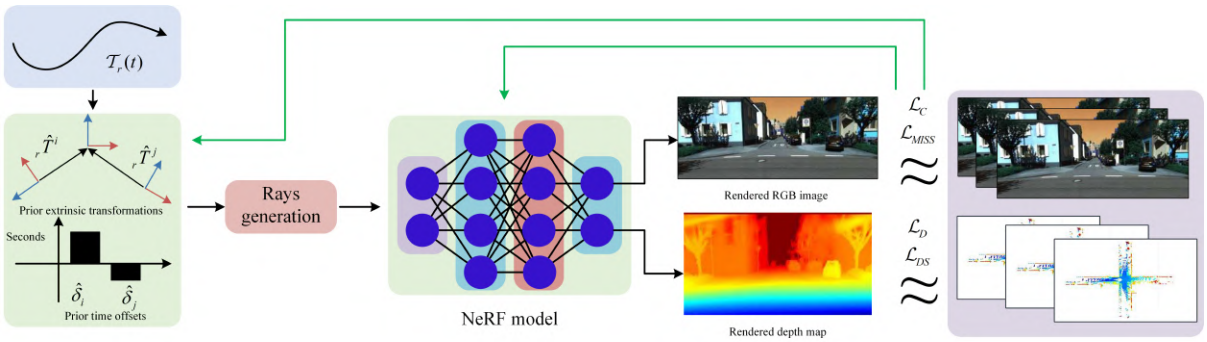

Figure 14: Framework of MOISST. Spatial extrinsic parameters and temporal offsets are optimized within a NeRF-based differentiable rendering loop using image, depth, and missing-observation consistency losses.

As illustrated in Fig. 14, MOISST converts spatial and temporal calibration errors into image and depth residuals within a shared implicit scene. This design extends calibration from a single LiDAR–camera pair to multi-sensor systems, where different cameras, LiDARs, and synchronization offsets can be constrained by the same scene representation.

Implicit radiance-field methods are attractive because they leverage dense image information, multi-frame observations, and continuous scene structure. However, their limitations are also evident: volume rendering is computationally expensive, training can be slow, and optimization remains sensitive to initialization, dynamic-object removal, ray sampling, and long-sequence drift.

6.2 3D Gaussian Splatting-Based Calibration

3D Gaussian Splatting represents a scene with explicit anisotropic Gaussians, each parameterized by position, covariance, color, and opacity [19]. In contrast to NeRF, 3DGS avoids dense volume sampling and renders images via differentiable rasterization. This results in higher efficiency for large-scale driving scenes and multi-frame calibration.

A 3D Gaussian scene can be written as

$$\mathcal{S}_{\text{3DGS}} = \left\{ \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \boldsymbol{c}_m, \alpha_m \right\}_{m=1}^{M}, \tag{21}$$

where $\boldsymbol{\mu}_m$ is the Gaussian center, $\boldsymbol{\Sigma}_m$ is the covariance, $\boldsymbol{c}_m$ is the color, and α_m is the opacity.

3DGS-Calib applies this representation to multimodal spatiotemporal calibration [24]. It initializes or constrains Gaussian positions with LiDAR geometry and jointly optimizes camera–LiDAR extrinsics together with Gaussian appearance and geometry. Compared with NeRF-based methods, it replaces slow volume rendering with faster Gaussian rasterization, making joint calibration more practical for vehicle-mounted multi-frame data.

However, 3DGS-based calibration introduces a coupling problem. Accurate extrinsics require an accurate scene model, while the scene model itself depends on correct extrinsics. If the initial calibration is poor, Gaussian positions, colors, scales, or opacities may partially absorb pose errors. Recent work therefore focuses on improving optimization stability. Constrained Gaussian Splatting investigates how pose and geometry constraints can prevent Gaussian deformation under inaccurate camera poses [92]. Self-calibrating Gaussian Splatting jointly optimizes camera parameters, lens distortion, and Gaussian fields, demonstrating that Gaussian representations can support broader calibration variables [93]. HiGS-Calib introduces a locally consistent photometric-geometric error model and a hierarchical coarse-to-fine optimization strategy to decouple scene modeling from extrinsic refinement. Targetless calibration with anchored 3D Gaussians employs reliable LiDAR points as anchor Gaussians to preserve scene scale and geometric stability [94].

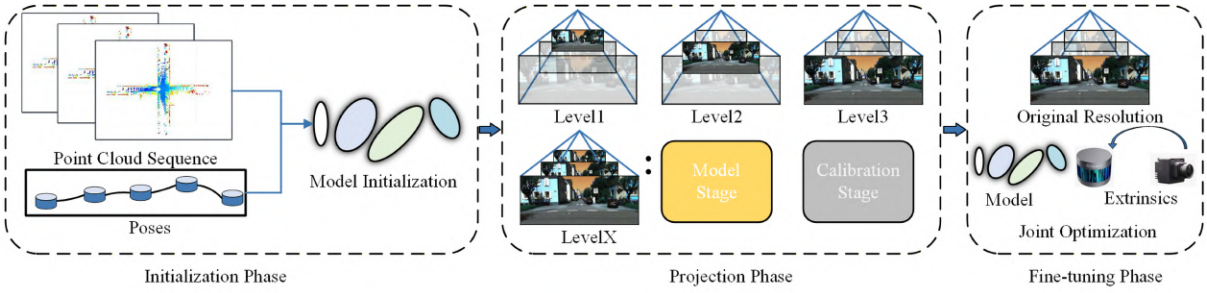

Figure 15: Framework of HiGS-Calib. The method initializes a Gaussian model from point-cloud sequences and poses, then performs hierarchical projection and fine-tuning to jointly optimize the scene model and LiDAR–camera extrinsics.

As illustrated in Fig. 15, hierarchical optimization separates initialization, multi-resolution projection, and final joint fine-tuning. This strategy helps reduce oscillation and local optima during Gaussian-based calibration by progressively improving both the scene representation and the extrinsic parameters.

Engineering deployment of 3DGS-based calibration may benefit from recent Gaussian SLAM systems. Gaussian Splatting SLAM and G-SLAM demonstrate the potential of Gaussian fields for dense mapping and real-time visual SLAM [95, 96]. These systems suggest that future calibration methods could be embedded into incremental mapping and local submap maintenance. Current challenges include memory cost, coupling of variables, dynamic-scene filtering, and interpretable error diagnosis.

6.3 2D Gaussian Splatting and Geometry-Enhanced Calibration

2D Gaussian Splatting is more oriented toward surface representation. It represents local surface patches and supports normals, depth continuity, and multi-view geometric consistency [20]. For LiDAR–camera calibration, this is advantageous because LiDAR naturally provides sparse depth and surface geometry, while camera images furnish texture and color.

Robust LiDAR–camera calibration with 2D Gaussian Splatting first reconstructs a color-free 2DGS geometry from LiDAR observations, then updates Gaussian colors using camera images while optimizing extrinsics. This design decouples the roles of the two modalities: LiDAR stabilizes geometry, while cameras constrain appearance. To mitigate the limitations of pure photometric loss under shadows, weak texture, and occlusion, 2DGS-based calibration further incorporates reprojection, triangulation, and depth-weighted uncertainty terms.

A representative loss can be summarized as

$$\mathcal{L}_{\text{2DGS}} = \mathcal{L}_{\text{photo}} + \lambda_p \mathcal{L}_{\text{repr}} + \lambda_t \mathcal{L}_{\text{tri}} + \lambda_u \mathcal{L}_{\text{unc}}, \tag{22}$$

where $\mathcal{L}_{\text{repr}}$ denotes reprojection consistency, $\mathcal{L}_{\text{tri}}$ denotes triangulation or multi-view geometric consistency, and $\mathcal{L}_{\text{unc}}$ weights unreliable depth or surface estimates.

Compared with 3DGS, 2DGS offers stronger geometric interpretability because its representation more closely approximates local surfaces. However, its reliability still depends on point-cloud density, surface-normal estimation, trajectory quality, and the static-scene assumption. If distant LiDAR points are too sparse, dynamic objects are frequent, or large planar regions lack texture, the resulting surface constraints may become weak.

6.4 Discussion

Differentiable scene representation-based calibration extends learning-based calibration from correspondence learning to scene-level joint optimization. Instead of estimating extrinsics from a limited set of local matches, these methods leverage rendered images, rendered depths, surface geometry, visibility, and regularization to constrain calibration. Implicit radiance fields emphasize continuous representation and unified scene explanation. 3DGS prioritizes efficiency and explicit structure. 2DGS highlights surface geometry and interpretable constraints.

These methods do not replace classical geometry or learning-based matching. Rather, they provide a higher-level optimization framework in which geometric constraints, learned priors, and rendering consistency can be integrated. Future research is likely to focus on lightweight online optimization, dynamic-scene modeling, uncertainty-aware calibration, multi-sensor spatiotemporal calibration, and integration with SLAM or mapping systems.

7 Challenges and Future Directions

The preceding sections have systematically reviewed four typical paradigms of LiDAR–camera extrinsic calibration from the perspectives of constraint construction and optimization strategies, each exhibiting distinct characteristics in applicability and performance. Table 5 summarizes the comprehensive properties of these four categories, covering typical representatives, validation datasets, achievable accuracy, initialization requirements, online calibration support, runtime efficiency, applicable scenarios, and primary advantages and limitations.

Table 5: Comprehensive comparison of four categories of LiDAR–camera calibration methods.

Method Category	**Target-Based**	**Targetless Explicit Matching**	**Learning-Based**	**Differentiable Scene Representation**
Typical Representatives	FAST-Calib	MLCC [10], Calib-Anything [75]	LCCNet [13], CFNet [14]	INF [21], 3DGS-Calib [24]
Validation Datasets	Custom calibration datasets	KITTI-360, WAYMO	KITTI, nuScenes	KITTI-360, nuScenes, Fast-LIVO2
Typical Accuracy (Rot./Trans.)	0.01°–0.05° / 1–3 mm	0.05°–0.2° / 3–10 mm	0.1°–0.5° / 5–16 mm	0.02°–0.1° / 2–8 mm
Initialization Requirement	Calibration target, multi-frame sequence	Rough initial value or scene structure	Training data, initial deviation range	Rough initial value, multi-frame sequence
Online Calibration Support	No	Partially supported	Yes	Yes (incremental optimization)
Time per Frame	< 1 s	0.2–2 s	50–200 ms	5–30 s (multi-frame)
Applicable Scenarios	Indoor, factory calibration	Structured urban scenes	Autonomous driving, specific scenarios	Multi-sensor mapping, complex scenes
Core Advantages	Highest accuracy, high reproducibility	No target required, strong physical interpretability	High automation, fast inference	Global consistency, handles complex scenes
Main Limitations	Manual intervention required, not online	Scene-dependent, prone to degeneracy	Limited generalization, weak interpretability	Computationally expensive, coupling issues

As shown in Table 5, no single paradigm consistently outperforms others across all metrics. Target-based methods achieve the highest accuracy yet remain constrained by manual deployment. Targetless explicit matching eliminates the need for calibration targets while remaining susceptible to scene degeneracy. Learning-based methods offer rapid inference and high automation, but their generalization capability is limited by domain shift. Differentiable scene representation methods provide global consistency at the cost of substantial computational overhead and potential coupling between scene reconstruction and extrinsic estimation.

Quantitative results on public datasets further reveal the accuracy degradation patterns under specific degenerate conditions. For planar-target-based methods with a single plane and limited viewpoints, the observability of rotation around the plane normal deteriorates severely, with rotation errors increasing by 0.5° to 1.2° compared with multi-view acquisitions. For targetless edge-alignment methods in weakly textured scenes, such as plain walls or flat roads, the availability of matching features between image edges and point cloud boundaries decreases substantially, leading to translation errors 5 cm to 12 cm higher than those in texture-rich scenes. For differentiable rendering methods based on Gaussian splatting or NeRF, when the initial extrinsic error exceeds 5° or 15 cm, the scene representation tends to absorb pose deviations to compensate for rendering losses, introducing an additional

rotation error of 0.6° to 1.5°. For deep learning-based end-to-end methods, domain shift between training and test scenarios results in a 100% to 180% increase in both rotation and translation errors, indicating a marked decline in generalization capability. These quantitative results demonstrate that each of the four categories has distinct performance boundaries and applicable conditions, and the choice of calibration method should be guided by the specific environmental constraints and accuracy requirements of the target application.

Despite the comparative landscape outlined above, no single paradigm offers universally reliable performance across all deployment scenarios. Target-based methods are accurate and reproducible, yet they require manual deployment and are unsuitable for continuous calibration after sensor installation. Targetless explicit matching improves practicality, but its performance hinges on the availability of stable edges, planes, semantic regions, or statistical correlations. Learning-based methods enhance automation and online correction capabilities, yet they demand representative training data and remain susceptible to domain shift. Differentiable scene representation methods provide a promising scene-level formulation, albeit at the cost of high computational expense and increased coupling between scene reconstruction and extrinsic estimation.

One persistent challenge is observability. Calibration is well constrained only when the scene, motion, and sensor overlap provide sufficient information for all six degrees of freedom. Single planar structures, repeated facades, weak texture, sparse distant LiDAR points, and narrow fields of view can all lead to degenerate or weakly constrained solutions. Future methods should therefore make observability explicit by estimating uncertainty, detecting degenerate configurations, and reporting when a calibration result is unreliable rather than merely returning a pose estimate.

Robustness in open environments is another central issue. Dynamic objects, occlusion, illumination changes, weather, rolling shutter, LiDAR motion distortion, and temporal offsets can introduce residuals that are unrelated to extrinsic error. Classical robust kernels, semantic filtering, and outlier rejection partially mitigate this problem, but online systems require more principled handling of uncertainty and temporal consistency. A promising direction is to combine geometry-based residuals, learned semantic priors, and differentiable rendering losses within an incremental framework that can update calibration while preserving long-term stability.

It is worth noting that a fundamental reason behind the difficulties faced by current LiDAR-camera calibration and broader scene representation tasks lies in the inherent mismatch of information dimensions: conventional LiDAR captures only 3D spatial information, whereas cameras capture 2D spatial plus spectral information. Hyperspectral LiDAR offers a fundamentally new solution to this problem by simultaneously acquiring 3D spatial coordinates and spectral reflectance properties of targets, thereby naturally bridging the modality gap at the data level. The acquired data can be regarded as native fusion of point clouds and hyperspectral images rather than post-hoc alignment, with both modalities already spatiotemporally aligned within the same sensor. Under this

condition, the traditional extrinsic calibration problem is largely dissolved, and the calibration task shifts from pose estimation between heterogeneous sensors to self-consistency verification of homogeneous multi-dimensional data.

Recent breakthroughs have also been achieved in LiDAR hardware and optical manipulation. Gong et al. proposed a novel LiDAR architecture based on an astigmatic metalens, achieving a point cloud acquisition rate of 36.6 MHz and a 102° field of view through spectral-acoustic cooperative scanning, providing a new hardware foundation for high-precision 3D perception [97]. Under extreme conditions, single-pixel single-photon LiDAR has demonstrated rapid 3D scanning imaging in low signal-to-noise ratio environments [98]. Furthermore, cross-scale vectorial optical field manipulation has opened new perspectives for surpassing the diffraction limit in long-range laser engineering [99]. Although hyperspectral LiDAR still faces engineering challenges in cost and form factor, its implications are worth attention. When a single sensor can simultaneously acquire spatial and spectral information, whether the classical cross-sensor extrinsic calibration problem remains a necessary step deserves reconsideration. Hyperspectral LiDAR offers new perspectives for calibration research from both hardware and algorithmic aspects: the former holds the potential to fundamentally reduce reliance on extrinsic parameter estimation accuracy, while the latter can provide a natural cross-modal association bridge. This direction encourages researchers to rethink the future paradigm of cross-modal information fusion.

Beyond hardware-level innovations, methodological insights from the broader multi-sensor fusion community also merit attention. A recent survey on resilient fusion navigation based on multisource PNT information [100] offers several concepts transferable to LiDAR–camera calibration. Multi-source redundancy enables verification using additional cues (e.g., GNSS/INS trajectories) beyond the sensor pair, resilience suggests graceful degradation under challenging conditions rather than blind pose output, and fault-detection frameworks can inspire outlier-aware correspondence rejection. This system-level perspective encourages treating calibration as an integrated component within a broader perception architecture, which is particularly relevant for long-term autonomous systems operating in dynamic and uncertain environments.

The recent rise of NeRF, 3DGS, and 2DGS suggests that future calibration may become part of mapping, localization, and scene reconstruction rather than an isolated preprocessing step. However, scene-level optimization must avoid allowing the representation to absorb calibration errors through deformation, opacity changes, or appearance compensation. Anchored geometry, multi-resolution optimization, physical sensor models, and uncertainty-aware regularization are therefore essential for making differentiable scene representation-based calibration reliable in practice.

Finally, fair evaluation remains difficult. Existing studies use different datasets, sensor configurations, initialization ranges, and error metrics, which complicates direct comparison. More standardized benchmarks should

evaluate not only final rotation and translation errors, but also convergence range, robustness to dynamic scenes, temporal offset sensitivity, computational cost, and failure detection. Such evaluation would facilitate comparisons among target-based, targetless, learning-based, and differentiable representation-based methods under realistic deployment conditions.

8 Conclusions

This review has examined LiDAR–camera extrinsic calibration through the lens of cross-modal evidence construction. Target-based methods create reliable shared observations via artificial geometry and remain the most robust solution for offline calibration. Targetless methods eliminate dedicated calibration objects by exploiting natural-scene structures, but their reliability depends on scene content and sensor overlap. Learning-based methods enhance automation by learning correspondences, calibration flow, or extrinsic corrections from data, yet remain challenged in generalization and interpretability. Differentiable scene representation-based methods further shift calibration toward joint scene-level optimization, where images, point clouds, rendering consistency, and sensor parameters are optimized together.

Across these paradigms, the central problem persists: heterogeneous LiDAR and camera measurements must be converted into constraints that are informative, robust, and observable. The field is thus moving toward hybrid calibration systems that combine explicit geometric constraints, learned cross-modal priors, uncertainty estimation, and efficient differentiable scene representations. Such systems are expected to enable more reliable online calibration for autonomous driving, robotics, mapping, and other long-term multi-sensor perception applications.

References

1. Li X, Xiao Y, Wang B et al. Automatic targetless LiDAR-camera calibration: A survey. *Artificial Intelligence Review* **56**, 9949–9987 (2023).
2. Liu Z, Chen Z, Wei X et al. External extrinsic calibration of multi-modal imaging sensors: A review. *IEEE Access* **11**, 110417–110441 (2023).
3. An P, Ding J, Quan S et al. Survey of extrinsic calibration on LiDAR-camera system for intelligent vehicle: Challenges, approaches, and trends. *IEEE Transactions on Intelligent Transportation Systems* **25**, 15342–15366 (2024).
4. Tan Z, Zhang X, Teng S et al. A review of deep learning-based LiDAR and camera extrinsic calibration. *Sensors* **24**, 3878 (2024).

5. Zhang H, Li S, Zhu X et al. 3-d LiDAR and monocular camera calibration: A review. *IEEE Sensors Journal* **25**, 10530–10555 (2025).

6. Wang S, Zhang P, Wu Y. Review of extrinsic parameter calibration of LiDAR and camera. *Virtual Reality & Intelligent Hardware* **8**, 28 (2026).

7. Scaramuzza D, Harati A, Siegwart R. Extrinsic self calibration of a camera and a 3d laser range finder from natural scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4164–4169 (2007).

8. Zhou L, Li Z, Kaess M. Automatic extrinsic calibration of a camera and a 3d LiDAR using line and plane correspondences. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5562–5569 (2018).

9. Yuan C, Liu X, Hong X et al. Pixel-level extrinsic self calibration of high resolution LiDAR and camera in targetless environments. *IEEE Robotics and Automation Letters* **6**, 7517–7524 (2021).

10. Liu X, Yuan C, Zhang F. Targetless extrinsic calibration of multiple small FoV LiDARs and cameras using adaptive voxelization. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–12 (2022). Art. no. 8502612.

11. Pandey G, McBride JR, Savarese S et al. Automatic extrinsic calibration of vision and LiDAR by maximizing mutual information. *Journal of Field Robotics* **32**, 696–722 (2015).

12. Iyer G, Ram RK, Murthy JK et al. CalibNet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1110–1117 (2018).

13. Lv X, Wang B, Dou Z et al. LCCNet: LiDAR and camera self-calibration using cost volume network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2894–2901 (2021).

14. Lv X, Wang S, Ye D. CFNet: LiDAR-camera registration using calibration flow network. *Sensors* **21**, 8112 (2021).

15. Jing X, Ding X, Xiong R et al. DXQ-Net: Differentiable LiDAR-camera extrinsic calibration using quality-aware flow. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 6235–6241 (2022).

16. Yuan K, Guo Z, Wang ZJ. RGGNet: Tolerance aware LiDAR-camera online calibration with geometric deep learning and generative model. *IEEE Robotics and Automation Letters* **5**, 6956–6963 (2020).

17. Mildenhall B, Srinivasan PP, Tancik M et al. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, 405–421 (2020).

18. Lin CH, Ma WC, Torralba A et al. BARF: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5741–5751 (2021).

19. Kerbl B, Kopanas G, Leimkühler T et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* **42**, 1–14 (2023).

20. Huang B, Yu Z, Chen A et al. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH Conference Papers*, 1–11 (2024).

21. Zhou S, Xie S, Ishikawa R et al. INF: Implicit neural fusion for LiDAR and camera. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 10918–10925 (2023).

22. Herau Q, Piasco N, Bennehar M et al. MOISST: Multimodal optimization of implicit scene for spatiotemporal calibration. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1810–1817 (2023).

23. Herau Q, Piasco N, Bennehar M et al. SOAC: Spatio-temporal overlap-aware multi-sensor calibration using neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15131–15140 (2024).

24. Herau Q, Bennehar M, Moreau A et al. 3DGS-Calib: 3d gaussian splatting for multimodal spatiotemporal calibration. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 8315–8321 (2024).

25. Zhang Q, Pless R. Extrinsic calibration of a camera and laser range finder. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* **3**, 2301–2306 (2004).

26. Geiger A, Moosmann F, Car Ö et al. Automatic camera and range sensor calibration using a single shot. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 3936–3943 (2012).

27. Beltrán J, Guindel C, de la Escalera A et al. Automatic extrinsic calibration method for LiDAR and camera sensor setups. *IEEE Transactions on Intelligent Transportation Systems* **23**, 17677–17689 (2022).

28. Yan G, He F, Shi C et al. Joint camera intrinsic and LiDAR-camera extrinsic calibration. In *2023 IEEE International Conference on Robotics and Automation*, 11446–11452 (2023).

29. Gong X, Lin Y, Liu J. 3d LIDAR-camera extrinsic calibration using an arbitrary trihedron. *Sensors* **13**, 1902–1918 (2013).

30. Kümmerle J, Kühner T, Lauer M. Automatic calibration of multiple cameras and depth sensors with a spherical target. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5584–5591 (2018).

31. Tóth T, Pusztai Z, Hajder L. Automatic LiDAR-camera calibration of extrinsic parameters using a spherical

target. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 8580–8586 (2020).

32. Wang W, Sakurada K, Kawaguchi N. Reflectance intensity assisted automatic and accurate extrinsic calibration of 3d LiDAR and panoramic camera using a printed chessboard. *Remote Sensing* **9**, 851 (2017).

33. Zheng C, Zhang F. FAST-Calib: LiDAR-camera extrinsic calibration in one second. arXiv preprint arXiv: 2507.17210 (2025).

34. Gentilini L, Serio P, Donzella V et al. A target-based multi-LiDAR multi-camera extrinsic calibration system. arXiv preprint arXiv: 2507.16621 (2025).

35. Cattaneo D, Valada A. CMRNext: Camera to LiDAR matching in the wild for localization and extrinsic calibration. *IEEE Transactions on Robotics* **41**, 1995–2013 (2025).

36. Zhao Y, Zhao L, Yang F et al. Extrinsic calibration of high resolution LiDAR and camera based on vanishing points. *IEEE Transactions on Instrumentation and Measurement* **73**, 1–12 (2024).

37. Song W, Oh M, Lee J et al. Galibr: Targetless LiDAR-camera extrinsic calibration method via ground plane initialization. In *IEEE Intelligent Vehicles Symposium*, 217–223 (2024).

38. Zhang Y, Xu J, Ren W. PLK-Calib: Single-shot and target-less LiDAR-camera extrinsic calibration using plücker lines. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 16091–16097 (2025).

39. Zeng T, He D, Yan F et al. YOCO: You only calibrate once for accurate extrinsic parameter in LiDAR-camera systems. *Measurement Science and Technology* **36**, 075009 (2025).

40. Ye T, Xu W, Zheng C et al. MFCalib: Single-shot and automatic extrinsic calibration for LiDAR and camera in targetless environments based on multi-feature edge. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 864–871 (2024).

41. Moghadam P, Bosse M, Zlot R. Line-based extrinsic calibration of range and image sensors. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 3685–3691 (2013).

42. Liao Q, Liu M. Extrinsic calibration of 3d range finder and camera without auxiliary object or human intervention. In *Proceedings of the IEEE International Conference on Real-Time Computing and Robotics*, 42–47 (2019).

43. Zhang X, Zhu S, Guo S et al. Line-based automatic extrinsic calibration of LiDAR and camera. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 9347–9353 (2021).

44. Chen F, Li L, Zhang S et al. PBACalib: Targetless extrinsic calibration for high-resolution LiDAR-camera system based on plane-constrained bundle adjustment. *IEEE Robotics and Automation Letters* **8**, 304–311 (2023).

45. Wang W, Nobuhara S, Nakamura R et al. SOIC: Semantic online initialization and calibration for LiDAR and camera. arXiv preprint arXiv: 2003.04260 (2020).

46. Sun Y, Li J, Wang Y et al. ATOP: An attention-to-optimization approach for automatic LiDAR-camera calibration via cross-modal object matching. *IEEE Transactions on Intelligent Vehicles* **8**, 696–708 (2023).

47. Liao Y, Li J, Kang S et al. SE-Calib: Semantic edge-based LiDAR-camera boresight online calibration in urban scenes. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 1–13 (2023).

48. Dhall A, Chelani K, Radhakrishnan V et al. LiDAR-camera calibration using 3d–3d point correspondences. arXiv preprint arXiv: 1705.09785 (2017).

49. Wang L, Xiao Z, Zhao D et al. Automatic extrinsic calibration of monocular camera and LiDAR in natural scenes. In *Proceedings of the IEEE International Conference on Information and Automation*, 997–1002 (2018).

50. Li J, Yang B, Chen C et al. Automatic registration of panoramic image sequence and mobile laser scanning data using semantic features. *ISPRS Journal of Photogrammetry and Remote Sensing* **136**, 41–57 (2018).

51. Nagy B, Benedek C. On-the-fly camera and LiDAR calibration. *Remote Sensing* **12**, 1137 (2020).

52. Hu H, Han F, Bieder F et al. TEScalib: Targetless extrinsic self-calibration of LiDAR and stereo camera for automated driving vehicles with uncertainty analysis. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 6256–6263 (2022).

53. Song J, Lee J. Online self-calibration of 3d measurement sensors using a voxel-based network. *Sensors* **22**, 6447 (2022).

54. Daniilidis K. Hand-eye calibration using dual quaternions. *The International Journal of Robotics Research* **18**, 286–298 (1999).

55. Ishikawa R, Oishi T, Ikeuchi K. LiDAR and camera calibration using motions estimated by sensor fusion odometry. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 7342–7349 (2018).

56. Park C, Moghadam P, Kim S et al. Spatiotemporal camera-LiDAR calibration: A targetless and structureless approach. *IEEE Robotics and Automation Letters* **5**, 1556–1563 (2020).

57. Wang Y, Tong Y, Wang R et al. SPTL-LCC: Single-shot, pixel-level, target-free, and LiDAR-type agnostic LiDAR-camera extrinsic self-calibration. *IEEE Transactions on Aerospace and Electronic Systems* **61**, 17567–17583 (2025).

58. Castorena J, Kamilov US, Boufounos PT. Autocalibration of LiDAR and optical cameras via edge alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2862–2866

(2016).

59. Castorena J, Puskorius GV, Pandey G. Motion guided LiDAR-camera self-calibration and accelerated depth upsampling for autonomous vehicles. *Journal of Intelligent & Robotic Systems* **100**, 1129–1138 (2020).

60. Zhu W, Shan S, Tong C et al. Targetless automatic edge-to-edge extrinsic calibration of LiDAR-camera system based on pure laser reflectivity. *IEEE Transactions on Instrumentation and Measurement* **74**, 1–14 (2025).

61. Levinson J, Thrun S. Automatic online calibration of cameras and lasers. In *Robotics: Science and Systems*, Vol.2 No.7 (2013).

62. Kang J, Doh NL. Automatic targetless camera-LIDAR calibration by aligning edge with gaussian mixture model. *Journal of Field Robotics* **37**, 158–179 (2020).

63. Yin J, Yan F, Liu Y et al. Automatic and targetless LiDAR-camera extrinsic calibration using edge alignment. *IEEE Sensors Journal* **23**, 19871–19880 (2023).

64. Yu Y, Fan S, Li L et al. Automatic targetless monocular camera and LiDAR external parameter calibration method for mobile robots. *Remote Sensing* **15**, 5560 (2023).

65. Jeong HY, Son WH, Shin DW et al. Targetless LiDAR-camera extrinsic calibration via class-agnostic boundary mask alignment and SPSA-based optimization. *Sensors* **26**, 1501 (2026).

66. Taylor Z, Nieto J, Johnson D. Multi-modal sensor calibration using a gradient orientation measure. *Journal of Field Robotics* **32**, 675–695 (2015).

67. Irie K, Sugiyama M, Tomono M. Target-less camera-LiDAR extrinsic calibration using a bagged dependence estimator. In *Proceedings of the IEEE International Conference on Automation Science and Engineering*, 1340–1347 (2016).

68. Napier A, Corke P, Newman P. Cross-calibration of push-broom 2d LIDARs and cameras in natural scenes. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 3679–3684 (2013).

69. Ishikawa R, Zhou S, Sato Y et al. LiDAR-camera calibration using intensity variance cost. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 10688–10694 (2024).

70. Guislain M, Digne J, Chaine R et al. Fine scale image registration in large-scale urban LiDAR point sets. *Computer Vision and Image Understanding* **157**, 90–102 (2017).

71. Koide K, Oishi S, Yokozuka M et al. General, single-shot, target-less, and automatic LiDAR-camera extrinsic calibration toolbox. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 11301–11307 (2023).

72. Jiang P, Osteen PR, Saripalli S. SemCal: Semantic LiDAR-camera calibration using neural mutual informa-

tion estimator. In *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 1–7 (2021).

73. Kodaira A, Zhou Y, Zang P et al. SST-Calib: Simultaneous spatial-temporal parameter calibration between LiDAR and camera. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2896–2902 (2022).

74. Zhu Y, Li C, Zhang Y. Online camera-LiDAR calibration with sensor semantic information. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 4970–4976 (2020).

75. Luo Z, Yan G, Li Y. Calib-anything: Zero-training LiDAR-camera extrinsic calibration method using segment anything. arXiv preprint arXiv: 2306.02656 (2023).

76. Huang Z, Zhang Y, Chen Q et al. Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching. *IEEE Transactions on Intelligent Vehicles* **10**, 3531–3542 (2025).

77. Chen X, Sun B. Targetless LiDAR-camera extrinsic calibration via semantic distribution alignment. *Frontiers in Robotics and AI* **13**, 1760867 (2026).

78. Han L, Ji M, Fang L et al. RegNet: Learning the optimization of direct image-to-image pose registration. arXiv preprint arXiv: 1812.10212 (2018).

79. Wang G, Qiu J, Guo Y et al. FusionNet: Coarse-to-fine extrinsic calibration network of LiDAR and camera with hierarchical point-pixel fusion. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 8964–8970 (2022).

80. Wu S, Hadachi A, Vivet D et al. NetCalib: A novel approach for LiDAR-camera auto-calibration based on deep learning. In *Proceedings of the International Conference on Pattern Recognition*, 6648–6655 (2021).

81. Cattaneo D, Vaghi M, Ballardini AL et al. CMRNet: Camera to LiDAR-map registration. In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 1283–1289 (2019).

82. Li X, Duan Y, Wang B et al. EdgeCalib: Multi-frame weighted edge features for automatic targetless LiDAR-camera calibration. *IEEE Robotics and Automation Letters* **9**, 10073–10080 (2024).

83. Shi J, Zhu Z, Zhang J et al. CalibRCNN: Calibrating camera and LiDAR by recurrent convolutional neural network and geometric constraints. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 10197–10202 (2020).

84. Xu S, Zhou S, Tian Z et al. RobustCalib: Robust LiDAR-camera extrinsic calibration with consistency learning. arXiv preprint arXiv: 2312.01085 (2023).

85. Xiao Y, Li Y, Meng C et al. CalibFormer: A transformer-based automatic LiDAR-camera calibration network. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 16714–16720 (2024).

86. Tian R, Bao X, Chen Y et al. RLCFormer: Automatic roadside LiDAR-camera calibration framework with transformer. *Heliyon* **10** (2024).

87. Wang Y, Zuo Y, Tang Y et al. Unsupervised optical-sensor extrinsic calibration via Dual-Transformer alignment. *Sensors* **25**, 6944 (2025).

88. Hu L, Wei C, Xu Y et al. Simple-calib: An end-to-end network for simultaneous calibration of multi LiDAR-camera. In *Proceedings of the IEEE International Conference on Unmanned Systems*, 1206–1211 (2025).

89. Hu L, Wei C, Wang M et al. Multi-calib: A scalable LiDAR–camera calibration network for variable sensor configurations. *Sensors* **25**, 7321 (2025).

90. Dash AR, Battrawy R, Schuster R et al. LiREC-Net: A target-free and learning-based network for LiDAR, rgb, and event calibration. arXiv preprint arXiv: 2602.21754 (2026).

91. Yang Z, Chen G, Zhang H et al. UniCal: Unified neural sensor calibration. In *European Conference on Computer Vision*, 327–345 (2024).

92. Peng J, Yan M, Huang S. A constrained optimization approach for gaussian splatting from coarsely-posed images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2961–2970 (2025).

93. Deng Y, Xian W, Yang G et al. Self-calibrating gaussian splatting for large field of view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 25124–25133 (2025).

94. Jung H, Kim N, Kim J et al. Targetless LiDAR-camera calibration with neural gaussian splatting. arXiv preprint arXiv: 2504.04597 (2025).

95. Matsuki H, Murai R, Kelly PHJ et al. Gaussian splatting SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18039–18048 (2024).

96. Yan C, Qu D, Wang D et al. GS-SLAM: Dense visual SLAM with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19595–19604 (2024).

97. Gong S, Guo Y, Li X et al. Spectral-acoustic-coordinated astigmatic metalens for wide field-of-view and high spatiotemporal resolution 3d imaging. *Light: Science & Applications* **15**, 85 (2026).

98. Jiang Y, Wang Z, Liu B et al. Fast three-dimensional scanning imaging using single pixel single photon lidar under extremely low signal to noise ratio conditions. *Intelligent Opto-Electronics* **1**, 250013–1–250013–12 (2025).

99. Guo Y, Pu M, Li Y et al. Surpassing the diffraction limit in long-range laser engineering via cross-scale vectorial optical field manipulation: perspectives and outlooks. *Opto-Electronic Advances* **8**, 250256–1–250256–10 (2025).

100. Zhao L, Fan Z, Yang C. Advances and challenges in resilient fusion navigation technology based on multi-source PNT information: A review. *IEEE Sensors Journal* **26**, 20608–20619 (2026).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 42274037, Grant 42574049, Grant 62271091, and Grant 62501088, by the Beijing Natural Science Foundation under Grant 4252042, and in part by the Fundamental Research Funds for the Central Universities under Grant 2024CDJXY020. The authors would like to thank the editors and reviewers for their valuable comments and suggestions.

Author contributions

Fengli Yang: Conceptualization, Investigation, Literature review, Writing – Original Draft (lead), Visualization, Formal analysis. **Jiangnan Peng**: Investigation, Literature review, Writing – Original Draft (supporting), Validation, Formal analysis, Writing – Review & Editing. **Zhihong Zeng**: Investigation, Literature review, Software (bibliographic analysis), Visualization, Writing – Review & Editing. **Dengke Wang**: Investigation, Literature review, Visualization, Writing – Review & Editing. **Chen Chen**: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – Review & Editing. **Long Zhao**: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – Review & Editing. **Harald Haas**: Supervision, Writing – Review & Editing, Validation, Resources. All authors have read and agreed to the published version of this manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

This article has no supplementary information.

Authors

Corresponding authors:

- Chen Chen, School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China. Email: c.chen@cqu.edu.cn
- Long Zhao, School of Space and Earth Sciences, Beihang University, Beijing 100191, China. Email: flylong@buaa.edu.cn

Co-authors:

- Fengli Yang, School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China.
- Jiangnan Peng, School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China.
- Zhihong Zeng, School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China.
- Dengke Wang, School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China.
- Harald Haas, Department of Engineering, Electrical Engineering Division, Cambridge University, CB3 0FA Cambridge, UK.