IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.
Cite as T-ASE paper.

Environment-Driven Online LiDAR-Camera Extrinsic Calibration

Zhiwei Huang, Jiaqi Li, Hongbo Zhao, Xiao Ma, Ping Zhong, *Member, IEEE*, Xiao-Hu Zhou, *Member, IEEE*, Wei Ye, *Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—LiDAR-camera extrinsic calibration (LCEC) is crucial for multi-modal data fusion in autonomous robotic systems. Existing methods, whether target-based or target-free, typically rely on customized calibration targets or fixed scene types, which limit their applicability in real-world scenarios. To address these challenges, we present EdO-LCEC, the first environment-driven online calibration approach. Unlike traditional target-free methods, EdO-LCEC employs a generalizable scene discriminator to estimate the feature density of the application environment. Guided by this feature density, EdO-LCEC extracts LiDAR intensity and depth features from varying perspectives to achieve higher calibration accuracy. To overcome the challenges of cross-modal feature matching between LiDAR and camera, we introduce dual-path correspondence matching (DPCM), which leverages both structural and textural consistency for reliable 3D-2D correspondences. Furthermore, we formulate the calibration process as a joint optimization problem that integrates global constraints across multiple views and scenes, thereby enhancing overall accuracy. Extensive experiments on real-world datasets demonstrate that EdO-LCEC outperforms state-of-the-art methods, particularly in scenarios involving sparse point clouds or partially overlapping sensor views.**

***Note to Practitioners*—This article presents an environment-driven approach for LiDAR-camera extrinsic calibration. Unlike conventional target-free methods, the proposed EdO-LCEC not only extracts matchable features from real-world scenes but also adapts its calibration strategy based on the environmental feature density. This environmental awareness significantly enhances calibration robustness. By focusing on cross-modal feature matching and extrinsic optimization, our method performs reliably across various sensor configurations, including solid-state and mechanical LiDARs with differing fields of view and point densities. The proposed approach offers a practical and generalizable solution that improves upon existing target-free methods, facilitating deployment in sensor fusion and mechatronic systems. Our calibration software developed on EdO-LCEC will be publicly available at https://mias.group/EdO-LCEC.**

***Index Terms*—Environment-driven, LiDAR-camera extrinsic calibration, multi-modal data fusion.**

Received 14 July 2025; revised 17 September 2025; accepted 11 October 2025. Date of publication 3 November 2025; date of current version 24 November 2025. This article was recommended for publication by Associate Editor Y. Sun and Editor P. Rocco upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62473288, Grant 62233013, Grant 62272489, and Grant 62388101; in part by the National Key Research and Development Program of China under Grant 2025YFE0200003; in part by the Fundamental Research Funds for the Central Universities; in part by Xiaomi Young Talents Program; and in part by the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, under Grant HMHAI-202406. *(Corresponding author: Rui Fan.)*

Zhiwei Huang, Jiaqi Li, Hongbo Zhao, and Wei Ye are with the Department of Control Science and Engineering, College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: 2431985@tongji.edu.cn; 2251550@tongji.edu.cn; hongbozhao@tongji.edu.cn; yew@tongji.edu.cn).

Xiao Ma is with Beijing Institute of Aerospace Control Devices, Beijing 100039, China (e-mail: mx_169@126.com).

Ping Zhong is with the School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China, and also with the National Key Laboratory of Science and Technology on Automatic Target Recognition, National University of Defense Technology, Changsha, Hunan 410073, China (e-mail: ping.zhong@csu.edu.cn).

Xiao-Hu Zhou is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xiaohu.zhou@ia.ac.cn).

Rui Fan is with the College of Electronic and Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai Key Laboratory of Intelligent Autonomous Systems, the State Key Laboratory of Autonomous Intelligent Unmanned Systems, and the Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Tongji University, Shanghai 201804, China, and also with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: rui.fan@ieee.org).

This article has supplementary downloadable material available at https://doi.org/10.1109/TASE.2025.3627253, provided by the authors.

Digital Object Identifier 10.1109/TASE.2025.3627253

I. Introduction

A. Background

PERCEPTION is a fundamental capability in autonomous mobile robotics, but achieving robust and generalizable perception remains a significant challenge [1]. By allowing robots to acquire and interpret information about their surroundings, it provides the critical foundation for dependable navigation, planning, and high-level decision-making [2]. Modern LiDAR-camera fusion systems significantly enhance robotic perception capabilities [3]. While LiDARs provide accurate spatial information, cameras capture rich textural details [4], [5]. Their complementary fusion enables robust performance in key robotic tasks, including SLAM [6], [7], [8], [9], [10], object recognition [11], [12], and localization [13]. LiDAR-camera extrinsic calibration (LCEC), which estimates the relative pose between the two sensors, becomes a core and foundational process for effective data fusion in automation science and engineering. Extensive research on offline, target-based LCEC has yielded numerous effective and robust algorithms over the decades. However, online target-free methods still struggle in complex, unstructured scenes due to limited adaptability to the working environment, and much work remains to fully realize their potential.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

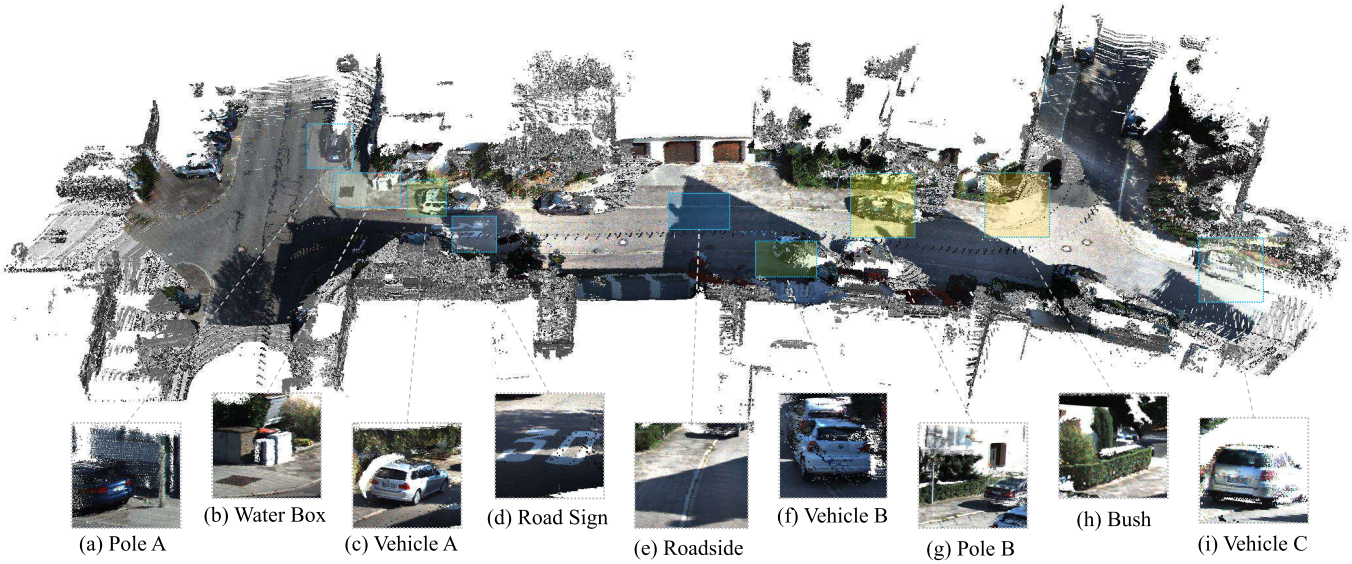

Fig. 1. Visualization of calibration results through LiDAR and camera data fusion in KITTI odometry 00 sequence: (a)-(i) zoomed-in regions that illustrate the alignment between the camera images and the LiDAR point clouds.

B. Existing Challenges and Motivation

Current LCEC methods are primarily categorized as either target-based or target-free. Target-based approaches [14], [15], [16], [17], [18] have long been the preferred choice in this field. They are typically offline, relying on customized calibration targets (typically checkerboards). However, they often demonstrate poor adaptability to real-world environments. This is largely because extrinsic parameters may change significantly due to moderate shocks or during extended operations in environments with vibrations.

Online, target-free approaches aim at overcoming this problem by extracting informative visual features directly from the environment. Previous works [19], [20] estimate the extrinsic parameters by matching the cross-modal edges between LiDAR projections and camera images. While effective in specific scenarios with abundant features, these traditional methods heavily rely on well-distributed edge features. Recent advances in deep learning techniques have spurred extensive explorations, such as [21], [22], and [23], to leverage semantic information to aid cross-modal feature matching. Although these approaches have shown compelling performance in specific scenarios, such as urban freeway, they predominantly rely on curated, pre-defined objects, *e.g.*, vehicles, lanes, and traffic poles. On the other hand, several end-to-end deep learning networks [24], [25], [26], [27], [28], [29] have been developed to find a more direct solution for LCEC. While these methods have demonstrated effectiveness on their training datasets, like KITTI [30], they highly rely on the training setup and are thus less generalizable. The recent work MIAS-LCEC [4] exploits the large vision model MobileSAM [31] to improve cross-modal feature matching. While this method achieves high accuracy with solid-state LiDARs (producing dense point clouds) when the sensors have overlapping fields of view, its performance significantly degrades in challenging scenarios with sparse or incomplete point clouds.

Although current target-free LCEC methods eliminate reliance on calibration targets, they are still restricted to specific sensor types and pre-defined environmental settings. As a result, they are hard to balance calibration accuracy and generalizability in complex, unstructured scenes due to limited adaptability. This challenge becomes even more pronounced in dynamic scenarios where sensor poses are uncertain and frequently changing. Therefore, this study aims to create a more flexible online calibration framework that adapts its strategy based on the environmental cues, thereby improving the robustness of online, target-free LCEC.

C. Novel Contributions

In this article, we move one step forward to introduce environmental observation into target-free calibration, proposing the first environment-driven framework, EdO-LCEC. EdO-LCEC adapts to external conditions to maintain optimal performance by balancing the feature density between LiDAR projection and camera image. Unlike the conventional target-free approaches, EdO-LCEC is no longer constrained by fixed strategies but can intelligently respond to environmental changes. Specifically, as illustrated in Fig. 1, we consider the working environment of the sensors as a sequence composed of multiple scenes. By actively perceiving the feature density of the environment and merging multiple scenes across different times, this approach could achieve high-precision calibration dynamically. A prerequisite for the success of EdO-LCEC is the design of a generalizable scene discriminator. The scene discriminator employs large vision models to conduct depth estimation and image segmentation. In detail, it calculates the feature density of the calibration scene and uses it to guide the generation of multiple virtual cameras for projecting LiDAR intensities and depth. This improved LiDAR point cloud projecting strategy increases the available environmental features, and thus overcomes the previous reliance

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

of algorithms [19], [32] on uniformly distributed geometric or textural features. At each scene, we perform dual-path correspondence matching (DPCM). Different from the C3M in MIAS-LCEC [4], DPCM divides correspondence matching into spatial and textural pathways to fully leverage both geometric and semantic information. Each pathway constructs a cost matrix based on structural and textural consistency, guided by accurate semantic priors, to yield reliable 3D-2D correspondences. Finally, the correspondences obtained from multiple views and scenes are used as inputs for our proposed multi-view and multi-scene joint optimization, which derives and refines the extrinsic matrix between LiDAR and camera. Through extensive experiments conducted on three real-world datasets, EdO-LCEC demonstrates superior robustness and accuracy compared to other SoTA approaches.

To summarize, our novel contributions are as follows:

- EdO-LCEC, the first environment-driven, online LCEC framework that introduces environmental observation into target-free calibration.
- Generalizable scene discriminator, which can automatically observe the calibration scene by evaluating the feature density through potential spatial, textural, and semantic features extracted by SoTA LVMs.
- DPCM, a novel cross-modal feature matching algorithm consisting of textural and spatial pathways, capable of generating dense and reliable 3D-2D correspondences between LiDAR point cloud and camera image.
- Multi-view and multi-scene joint relative pose optimization, enabling high-quality extrinsic estimation by integrating multiple perspective views within a single scene and merging distinct scenes across different timestamps.

D. Article Structure

The remainder of this article is structured as follows: Sect. II reviews SoTA approaches in LCEC. Sect. III introduces EdO-LCEC, our proposed online, target-free LCEC algorithm. Sect. IV presents experimental results and compares our method with SoTA methods. Finally, in Sect. V, we conclude this article and discuss potential future research directions.

II. Related Work

Target-based LCEC methods achieve high accuracy using customized calibration targets (typically checkerboards). However, they require offline execution and are significantly limited in dynamic or unstructured environments where such targets are unavailable [23], [33]. Recent studies have shifted to online, target-free approaches to overcome these limitations. Pioneering works [19], [20], [34], [35], [36] estimate the relative pose between the two sensors by aligning the cross-modal edges or mutual information (MI) extracted from LiDAR projections and camera RGB images. While effective in specific scenarios with abundant features, these traditional methods heavily rely on well-distributed edges and rich texture, which largely compromise calibration robustness. To circumvent the challenges associated with cross-modal feature matching, several studies [33], [37], [38], [39] have explored motion-based methods. These approaches match sensor motion trajectories from visual and LiDAR odometry to derive extrinsic parameters through optimization. While they effectively accommodate heterogeneous sensors without requiring overlap, they demand precise synchronized LiDAR point clouds and camera images to accurately estimate per-sensor motion, which limits their applicability in real-world scenarios.

Advances in deep learning techniques have driven significant exploration into enhancing traditional target-free algorithms. Some studies [4], [21], [22], [32], [40] explore attaching deep learning modules to their calibration framework as useful tools to enhance calibration efficiency. For instance, [40] accomplishes LiDAR and camera registration by aligning road lanes and poles detected by semantic segmentation. Similarly, [22] employs stop signs as calibration primitives and refines results over time using a Kalman filter. A recent study [32] introduced Direct Visual LiDAR Calibration (DVL), a novel point-based method that utilizes SuperGlue [41] to establish direct 3D-2D correspondences between LiDAR and camera data. On the other hand, several learning-based algorithms [25], [26], [27], [28], [29] attempt to reformulate the calibration process into more direct solutions by leveraging end-to-end deep learning networks. Although these methods have shown promising results on public datasets such as KITTI [30], which primarily focus on urban driving scenarios, their performance has not been extensively validated on other real-world datasets that include more diverse and challenging scenes. Moreover, end-to-end models trained on a single LiDAR-camera pair often overfit to the intrinsic and extrinsic parameters of the training setup and fail to generalize, even when evaluated on a different camera within the same dataset. For instance, a network trained end-to-end on the left camera of KITTI but evaluated on the right camera still predicts the pose of the left camera. This indicates that such networks primarily memorize specific extrinsic parameters rather than learning a generalizable calibration strategy. In contrast, our method incorporates feature density evaluation to achieve environment-aware calibration. This environment-driven strategy enables robots to maintain high-precision calibration across diverse conditions without being constrained by specific sensor configurations.

III. Methodology

Given LiDAR point clouds and camera images, our goal is to estimate their extrinsic matrix ${}^{C}_{L}\boldsymbol{T}$, defined as follows:

$$ {}^{C}_{L}\boldsymbol{T} = \begin{bmatrix} {}^{C}_{L}\boldsymbol{R} & {}^{C}_{L}\boldsymbol{t} \\ \boldsymbol{0}^{\top} & 1 \end{bmatrix} \in SE(3), \tag{1} $$

where ${}^{C}_{L}\boldsymbol{R} \in SO(3)$ represents the rotation matrix, ${}^{C}_{L}\boldsymbol{t}$ denotes the translation vector, and $\boldsymbol{0}$ represents a column vector of zeros. We first give an overview of the proposed method. As shown in Fig. 2, it mainly contains three stages:

- We first utilize a scene discriminator to perceive the environment through image segmentation and depth estimation, generating virtual cameras that project LiDAR intensity and depth from multiple viewpoints. LiDAR projections from multiple views are further segmented into masks with corner points (Sect. III-A).

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

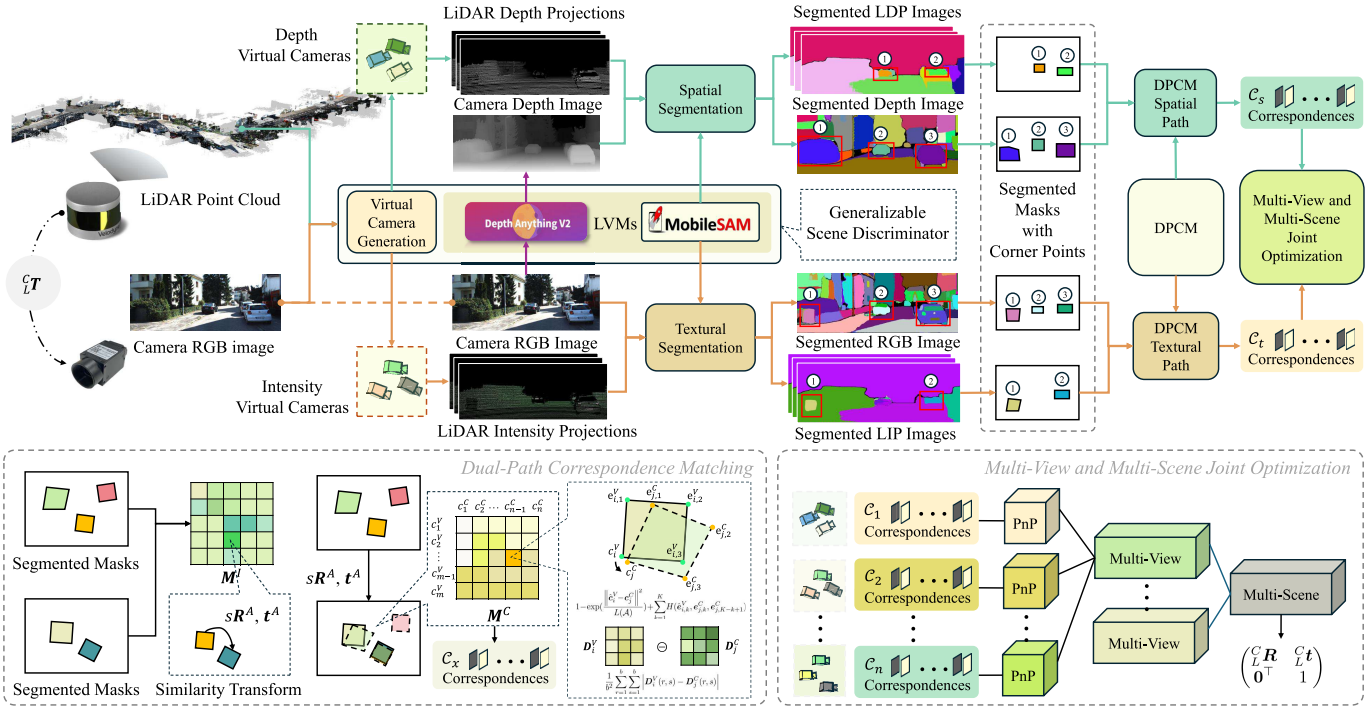

Fig. 2. The pipeline of our proposed EdO-LCEC. The working environment of the sensors is analyzed by the generalizable scene discriminator. In each calibration scene, the feature density of LiDAR and camera data is estimated by image segmentation and depth estimation. Based on this feature density, the scene discriminator generates multiple depth and intensity virtual cameras to create LIP and LDP images. Image segmentation results (segmented masks with corner points) of virtual images and camera images are sent to DPCM to obtain dense 3D-2D correspondences, which serve as input for the multi-view and multi-scene joint optimization to derive and refine the extrinsic matrix between LiDAR and camera.

- The segmented masks with detected corner points are processed along two pathways (spatial and textural) of the dual-path correspondence matching module to establish reliable 3D-2D correspondences (Sect. III-B).
- The obtained correspondences are used as inputs for our proposed multi-view and multi-scene joint optimization method, which derives and refines the extrinsic matrix ${}^{C}_{L}\boldsymbol{T}$ (Sect. III-C).

A. Generalizable Scene Discriminator

Our environment-driven approach first employs a generalizable scene discriminator to observe the surroundings by generating virtual cameras to project LiDAR point cloud intensities and depth. This discriminator configures both an intensity and a depth virtual camera from the LiDAR's perspective. This setup yields a LiDAR intensity projection (LIP) image ${}^{V}_{I}\boldsymbol{I} \in \mathbb{R}^{H\times W\times 1}$ (H and W represent the image height and width) and a LiDAR depth projection (LDP) image ${}^{V}_{D}\boldsymbol{I}$. To align with the LDP image, the input camera RGB image ${}^{C}_{I}\boldsymbol{I}$ is processed using Depth Anything V2 [42] to obtain estimated depth images ${}^{C}_{D}\boldsymbol{I}$.[1] To take advantage of semantic information, we utilize MobileSAM [43] as the image segmentation backbone. The series of n detected masks in an image is defined as $\{\mathcal{M}_1, \ldots, \mathcal{M}_n\}$. The corner points along the contours of masks detected are represented by $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{m_i}\}$, where m_i is the total corner points number in the i-th mask. An instance (bounding box), utilized to precisely fit around each mask, is centrally positioned at $\boldsymbol{o}$ and has a dimension of $h \times w$ pixels. To fully exploit the textural information, we evaluate the consistency of corner points using a texture matrix $\boldsymbol{D} \in \mathbb{R}^{b\times b}$, which encodes intensity values within a local $b\times b$ neighborhood. As depicted in Fig. 2, the neighboring vertices of a corner point $\boldsymbol{c}_i$ is defined as $\{\boldsymbol{e}_{i,1}, \ldots, \boldsymbol{e}_{i,K}\}$. These neighboring vertices are used to calculate structural consistency in dual-path correspondence matching.

For each virtual or camera image $\boldsymbol{I}$, the scene discriminator computes its feature density $\rho(\boldsymbol{I})$, providing critical cues for feature extraction and correspondence matching. The feature density $\rho(\boldsymbol{I})$ is defined as follows:

$$\rho(\boldsymbol{I}) = \underbrace{\left(\log\left(\sum_{i=1}^{n} m_i\right)^2\right)}_{\rho_t} \underbrace{\left(\sum_{i=1}^{n} \log \frac{\left|\bigcup_{j=1}^{n} \mathcal{M}_j\right|}{|\mathcal{M}_i|}\right)}_{\rho_s}, \tag{2}$$

where ρ_t denotes the textural density and ρ_s represents the structural density. The occlusion challenges caused by the different perspectives of LiDAR and the camera [4], [19], combined with limited feature availability in sparse point clouds, mean that a single pair of virtual cameras is insufficient for a comprehensive view of the calibration scene. Let E represent the event of capturing enough effective features, with probability $P(E) = \lambda$. If we have n_I intensity virtual cameras and n_D depth virtual cameras, the probability of capturing enough effective features is $1-(1-\lambda)^{n_I+n_D}$. In theory,

[1] In this article, the symbols in the superscript denote the type of target camera (V denotes virtual camera and C indicates real camera), and the subscript denotes the source of the image (D is depth and I is intensity).

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

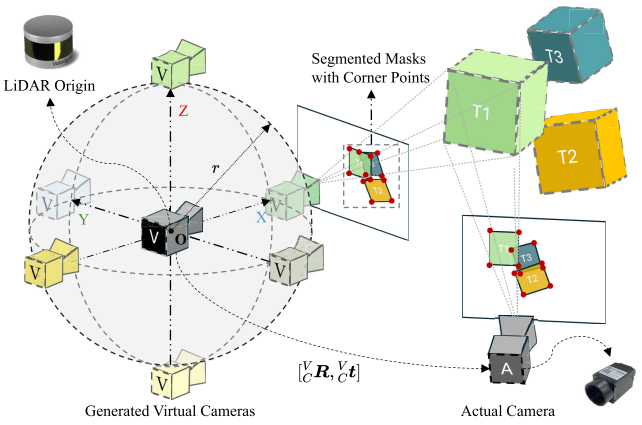

Fig. 3. Virtual camera generation method. We distribute the virtual cameras along the X, Y, and Z axes inside a sphere of radius $r = 0.3$ m. The scene discriminator dynamically determines the number of virtual cameras based on feature density. All cameras maintain a default front-facing orientation.

as $n_I + n_D \to \infty$, the probability $P(E')^{n_I+n_D} \to 0$, leading to $1-(1-\lambda)^{n_I+n_D} \to 1$. Increasing the number of virtual cameras raises the likelihood of capturing more potential correspondences, thus enhancing calibration accuracy.

However, employing an infinite number of virtual cameras during calibration is impractical. In real applications, a straightforward approach is to let users specify the number of virtual cameras based on their experience. If an initial estimate of the calibration parameters is available, the number of virtual cameras can also be inferred from the distance between the LiDAR origin and the initial estimate. Another automated option is to compute the ratio between the LiDAR and camera fields of view (FoVs) to determine the minimum number of virtual cameras required. Each virtual camera replicates the intrinsic FoV of the real camera, and together they fully cover the LiDAR scanning range. While the above strategies enhanced calibration performance in certain specific cases, they lacked generalizability across more complex and diverse scenarios. The required parameters vary with environmental changes, and the absence of environmental perception prevents the algorithm from automatically adapting to unseen and challenging conditions.

In this work, we introduce a fully automated strategy for generating virtual cameras guided by environmental information. Considering the trade-off between calibration accuracy and computational cost, we determine the number of virtual cameras to balance the feature density:

$$\rho({}^C_I\boldsymbol{I}) + \rho({}^C_D\boldsymbol{I}) = \sum_{i=0}^{n_I-1} \rho({}^V_I\boldsymbol{I}_i) + \sum_{i=0}^{n_D-1} \rho({}^V_D\boldsymbol{I}_i). \tag{3}$$

As depicted in Fig. 3, in practical applications, we set multiple virtual cameras inside a sphere originating from the LiDAR perspective center. A smaller or larger sphere might be used to set more cameras if necessary. Since the feature density is similar if the perspective viewpoints of virtual cameras are close to the initial position, we can assume that $\rho({}^V_I\boldsymbol{I}_i) \approx \rho({}^V_I\boldsymbol{I}_0)$ and $\rho({}^V_D\boldsymbol{I}_i) \approx \rho({}^V_D\boldsymbol{I}_0)$. So n_I and n_D can be obtained as follows:

$$n_I = \frac{\rho({}^C_I\boldsymbol{I})}{\rho({}^V_I\boldsymbol{I}_0)}, \quad n_D = \frac{\rho({}^C_D\boldsymbol{I})}{\rho({}^V_D\boldsymbol{I}_0)}. \tag{4}$$

Once all virtual cameras are generated, the discriminator performs image segmentation on each LiDAR projection captured from multiple views, detecting the corner points of the masks. These masks with detected corner points serve as inputs for the dual-path correspondence matching.

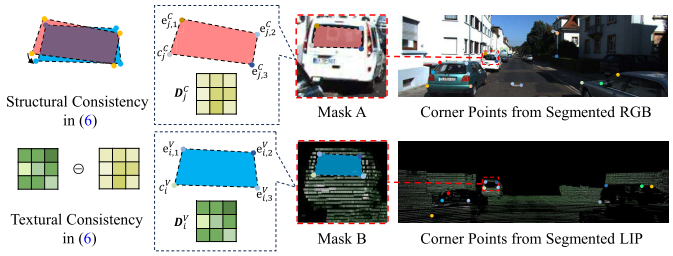

Fig. 4. DPCM utilizes structural consistency and textural consistency in (6) to compute matching cost between corner points detected from different segmented masks.

B. Dual-Path Correspondence Matching

Given the segmented masks with detected corner points, dual-path correspondence matching leverages them to achieve dense and reliable 3D-2D correspondences. DPCM consists of two pathways, one for correspondence matching of LIP and RGB images, and the other for LDP and depth images. For each pathway, DPCM adopted the approach outlined in [4] to obtain mask matching result $\mathcal{A} = \{(\mathcal{M}_i^V, \mathcal{M}_i^C) \mid i = 1, \ldots, m\}$ from a cost matrix $\boldsymbol{M}^I$. Each matched mask pair can estimate a 4-DoF similarity transform $[s\boldsymbol{R}^A, \boldsymbol{t}^A]$ to guide the correspondence matching. Specifically, we update the corner points $\boldsymbol{c}_i^V$ in the virtual image to a location $\hat{\boldsymbol{c}}_i^V$ that is close to its true projection coordinate using this affine transformation, as follows:

$$\hat{\boldsymbol{c}}_i^V = s\boldsymbol{R}^A(\boldsymbol{c}_i^V) + \boldsymbol{t}^A. \tag{5}$$

To determine optimum corner point matches, we construct a cost matrix $\boldsymbol{M}^C$, where the element at $\boldsymbol{x} = [i, j]^\top$, namely:

$$\begin{aligned}\boldsymbol{M}^C(\boldsymbol{x}) = &\\ &\underbrace{\beta_s\left(1-\exp\left(\frac{\left\|\hat{\boldsymbol{c}}_i^V-\boldsymbol{c}_j^C\right\|^2}{L(\mathcal{A})}\right)+\sum_{k=1}^{K} H(\hat{\boldsymbol{e}}_{i,k}^V, \boldsymbol{e}_{j,k}^C, \boldsymbol{e}_{j,K-k+1}^C)\right)}_{\text{Structural Consistency}}\\ &+\underbrace{\beta_t\left(\frac{1}{b^2}\sum_{r=1}^{b}\sum_{s=1}^{b}\left|\boldsymbol{D}_i^V(r,s)-\boldsymbol{D}_j^C(r,s)\right|\right)}_{\text{Textural Consistency}}\end{aligned} \tag{6}$$

denotes the matching cost between the i-th corner point of a mask in the LiDAR virtual image and the j-th corner point of a mask in the camera image. (6) consists of structural and textural consistency. As illustrated in Fig. 4, the structural consistency measures the structural difference of corner points in the virtual and real image, where $L(\mathcal{A})$ serves as a width

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

Algorithm 1 Dual-Path Correspondence Matching

Require:
Textural pathway: Segmented Masks (including detected corner points) obtained from the LIP and RGB image.
Spatial pathway: Segmented Masks (including detected corner points) obtained from the LDP and depth images.
Stage 1 (Reliable mask matching):
(1) Conduct cross-modal mask matching by adopting the method described in [4].
(2) Estimate $s\boldsymbol{R}^A$ and $\boldsymbol{t}^A$ using the approach in [4].
Stage 2 (Dense correspondence matching):
(1) For each pathway, update all masks in the virtual image using $s\boldsymbol{R}^A$ and $\boldsymbol{t}^A$.
(2) Construct the corner point cost matrix $\boldsymbol{M}^C$ using (6).
(3) Select matches with the lowest costs in both horizontal and vertical directions of $\boldsymbol{M}^C(\boldsymbol{x})$ as the optimum correspondence matches.
(4) Aggregate all corner point correspondences to form the sets $\mathcal{C} = \{(\boldsymbol{p}_i^L, \boldsymbol{p}_i) \mid i = 1, \ldots, q\}$.

parameter based on the average perimeter of the matched masks and $H(\hat{\boldsymbol{e}}_{i,k}^V, \boldsymbol{e}_{j,k}^C, \boldsymbol{e}_{j,K-k+1}^C)$ represents the similarity of the neighboring vertices between current and target corner point. The textural consistency derives from the relative textural similarity of the b neighboring zone. After establishing the cost matrix, a strict criterion is applied to achieve reliable matching. Matches with the lowest costs in both horizontal and vertical directions of $\boldsymbol{M}^C(\boldsymbol{x})$ are determined as the optimum corner point matches. Since every $\boldsymbol{c}_i^V$ can trace back to a LiDAR 3D point $\boldsymbol{p}_i^L = [x^L, y^L, z^L]^\top$, and every $\boldsymbol{c}_i^C$ is related to a pixel $\boldsymbol{p}_i = [u, v]^\top$ (represented in homogeneous coordinates as $\tilde{\boldsymbol{p}}_i$) in the camera image, the final correspondence matching result of DPCM is $\mathcal{C} = \{(\boldsymbol{p}_i^L, \boldsymbol{p}_i) \mid i = 1, \ldots, q\}$.

Leveraging the generalizable scene discriminator, the multi-perspective views generated by virtual cameras provide rich spatial-textural descriptors that effectively transform the feature matching problem into a robust 3D-2D correspondence establishment process. DPCM achieves efficient computation by verifying 2D structural and textural consistency instead of performing costly 3D point cloud registration. The pseudo-code of our DPCM is presented in Algorithm 1. During corner point matching within a mask, while the C3M in MIAS-LCEC exclusively considers points from the corresponding mask (*i.e.*, the matched mask associated with the current corner point), our DPCM in EdO-LCEC incorporates additional corner points from neighboring masks. This innovative matching strategy offers two key advantages: (1) it partially eliminates the dependency on dense, high-precision mask matching results, and (2) it substantially avoids errors induced by imperfect mask matching, especially under challenging conditions such as sparse LiDAR point clouds or limited LiDAR-camera FoV overlap.

C. Multi-View and Multi-Scene Joint Optimization

EdO-LCEC models the sensor's operational environment as a composition of N distinct scenes across different timestamps. While single-view methods (*e.g.*, [4], [19]) suffer from limited high-quality correspondences in sparse or incomplete point cloud scenarios, our environment-driven approach overcomes this constraint through integrating observations from multiple views and scenes. After establishing 3D-2D correspondences by DPCM, EdO-LCEC computes the extrinsic matrix between LiDAR and camera through a multi-view and multi-scene joint optimization. By aggregating optimal correspondences across time and space, our method enhances matching robustness, maximizes high-quality feature associations, and ultimately improves calibration accuracy.

In multi-view optimization, the extrinsic matrix ${}_L^C\hat{\boldsymbol{T}}_t$ of the t-th scene can be formulated as follows:

$$ {}_L^C\hat{\boldsymbol{T}}_t = \arg\min_{{}_L^C\boldsymbol{T}_{t,k}} \sum_{i=1}^{n_I+n_D} \sum_{(\boldsymbol{p}_j^L, \boldsymbol{p}_j)\in\mathcal{C}_i} G\left(\underbrace{\left\| \pi({}_L^C\boldsymbol{T}_{t,k}\boldsymbol{p}_j^L) - \tilde{\boldsymbol{p}}_j \right\|_2}_{\epsilon_j} \right), \tag{7} $$

where ${}_L^C\boldsymbol{T}_{t,k}$ denotes the k-th PnP solution obtained using a selected subset $\mathcal{V}_k$ of correspondences from $\mathcal{C}_i$, and ϵ_j represents the reprojection error of $(\boldsymbol{p}_i^L, \boldsymbol{p}_i)$ with respect to ${}_L^C\boldsymbol{T}_{t,k}$. $G(\epsilon_j)$ represents the gaussian depth-normalized reprojection error under the projection model π, defined as:

$$ G(\epsilon_j) = \frac{\epsilon_j(e - \mathcal{K}(d_j', \bar{d}'))}{H + \epsilon_j}, \tag{8} $$

where d_j' is the normalized depth of $\boldsymbol{p}_j^L$, $\bar{d}'$ is the average normalized depth, and $\mathcal{K}(d_j', \bar{d}')$ is a gaussian kernal.

In multi-scene optimization, we choose a reliable subset $\mathcal{S}_t$ from each scenario, which can be obtained as follows:

$$ \mathcal{S}_t = \bigcup_{k=1}^{s_t} \mathcal{V}_{t,k}, \quad s_t = \min\left\{ \frac{Q_{\max} q_t}{\sum_{j=1}^N q_j}, s_{\max} \right\}, \tag{9} $$

where $\mathcal{V}_{t,k}$ is the k-th selected correspondences subset in the t-th scene. q_t denotes the number of correspondences in the t-th scene. $Q_{\max}$ represents the maximum number of correspondences hope to get from all scenarios, and $s_{\max}$ denotes the maximum number of reliable correspondences in a single scene. The multi-scene optimization process then solves the final extrinsic matrix ${}_L^C\boldsymbol{T}^*$ by minimizing the joint reprojection error:

$$ \mathcal{L}({}_L^C\boldsymbol{T}^*) = \sum_{t=1}^{N} \sum_{(\boldsymbol{p}_j^L, \boldsymbol{p}_j)\in\mathcal{S}_t} G\left(\left\| \pi({}_L^C\boldsymbol{T}^* \boldsymbol{p}_j^L) - \tilde{\boldsymbol{p}}_j \right\|_2 \right) \tag{10} $$

across multiple spaces in the environment at different times. This process combines optimal correspondences from both spatial and textural pathways in each scenario, enabling robust environment-driven optimization through multi-view and multi-scene fusion. As a result, our method achieves human-like adaptability to dynamic environments. This environment perception makes EdO-LCEC behave much better than other target-free approaches, especially in conditions when point clouds are sparse or incomplete.

IV. Experiment

A. Experimental Setup and Evaluation Metrics

In our experiments, we evaluate our proposed EdO-LCEC on four public datasets: KITTI odometry [30] (including

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

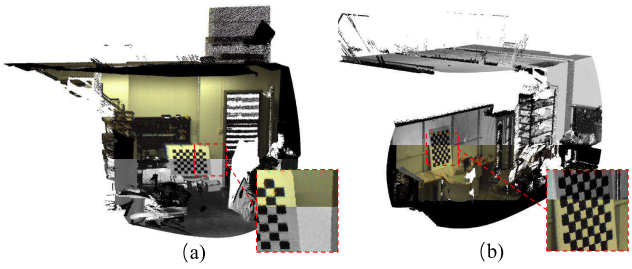

Fig. 5. Visualization of EdO-LCEC calibration results through LiDAR and camera data fusion: (a)-(b) illustrate two LiDAR point clouds in MIAS-LCEC-TF70, partially rendered by the image color using the estimated extrinsic matrix of EdO-LCEC.

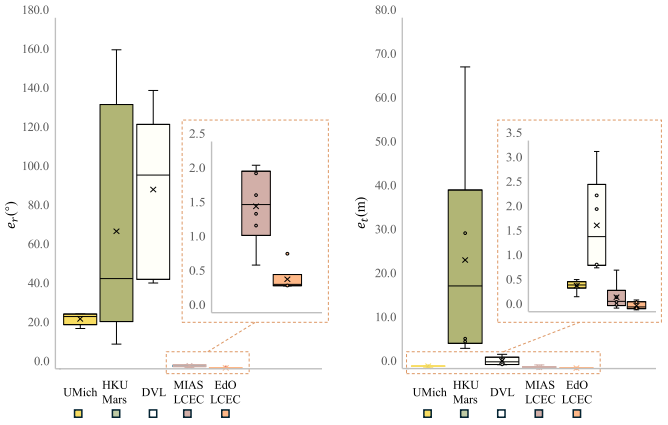

Fig. 6. Comparisons with SoTA approaches on the segmented point clouds from MIAS-LCEC-TF360. The zoomed-in region highlights the comparative details between the algorithms with higher accuracy. Since the results of CRLF are invalid, they were not included in the comparison.

00-09 sequences), KITTI360 [45], nuScenes [46], and MIAS-LCEC [4] (including target-free datasets MIAS-LCEC-TF70 and MIAS-LCEC-TF360). Extensive comparisons with SoTA LCEC approaches and abundant ablation studies are conducted to comprehensively validate each component of the proposed algorithm. Following previous work [4], [25], [28], [32], we utilize the average magnitude e_r of Euler angle error and the average magnitude e_t of the translation error to quantify the calibration errors. Each axes of the Euler angle error and translation error are also provided to comprehensively demonstrate the calibration accuracy.

Notably, sequences in KITTI odometry, aside from 00, were included in the training datasets for the learning-based methods [25], [27], [29]. To ensure a fair comparison, we reproduced calibration results for both the left and right cameras on sequence 00 when the authors provided their code; otherwise, we used the reported results for the left camera from their papers. Since most learning-based methods lack APIs for custom data, our comparison with these methods is limited to the KITTI odometry 00 sequence. In nuScenes, the point clouds are captured by a 32-line spinning LiDAR (Velodyne HDL32E), which makes them extremely sparse for previous calibration approaches. To the best of our knowledge, few LCEC methods have been evaluated on this dataset. Following the official split, we use the scenes in v1.0-test of nuScenes to construct the evaluation data. For experiments on the MIAS-LCEC dataset, as the results for the compared methods [4], [19], [20], [32], [40] are reported in [4], we directly use the values presented in that paper.

Our algorithm was implemented on an Intel i7-14700K CPU and an NVIDIA RTX4070Ti Super GPU. The entire process of a single-view calibration, including scene discriminating, DPCM, and relative pose optimization, takes approximately 15 to 70 seconds.

B. Comparison With State-of-the-Art Method

In this section, quantitative comparisons with SoTA approaches on three datasets are presented in Fig. 6, Tables I,[2] II, III, and IV. Additionally, qualitative results are illustrated in Figs. 5 and 7.

[2]The reproduced results of LCCNet yield higher errors compared to those reported in their paper.

1) Evaluation on KITTI Odometry: The results shown in Table I and II suggest that, with the exception of sequences 01 and 04, our method achieves SoTA performance across the ten sequences (00-09) in KITTI odometry. Specifically, in the 00 sequence, EdO-LCEC reduces the e_r by around 35.2-99.8% and the e_t by 16.1-98.8% for the left camera, and reduces the e_r by around 46.9-99.7% and the e_t by 13.9-97.6% for the right camera. Additionally, according to Fig. 7, it can be observed that the point cloud of a single frame in KITTI is so sparse that the other approaches behave poorly. In contrast, our proposed method overcomes this difficulty and achieves high-quality data fusion through the calibration result. We attribute these performance improvements to our multi-view and multi-scene joint optimization. Merging the optimal matching results from multiple views and scenes maximizes the number of reliable correspondences and ultimately improves overall calibration accuracy.

2) Evaluation on MIAS-LCEC Dataset: Compared to the sparse point clouds in KITTI odometry, the point clouds in the MIAS-LCEC datasets are significantly denser, which facilitates feature matching and allows us to test the upper limits of the algorithm calibration accuracy. The results shown in Table III demonstrate that our method outperforms all other SoTA approaches on MIAS-LCEC-TF70. It can also be observed that our method dramatically outperforms CRLF, UMich, DVL, HKU-Mars, and is slightly better than MIAS-LCEC across the total six subsets. In challenging conditions that are under poor illumination and adverse weather, or when few geometric features are detectable, EdO-LCEC performs significantly better than all methods, particularly. This impressive performance can be attributed to the generalizable scene discriminator. The multiple virtual cameras generated by the scene discriminator provide a comprehensive perception of the calibration scene from both spatial and textural perspectives, which largely increases the possibility of capturing high-quality correspondences for the PnP solver. Furthermore, the data fusion results in Fig. 5, obtained using our optimized extrinsic matrix, visually demonstrate perfect alignment on the checkerboard. This highlights the high calibration accuracy

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

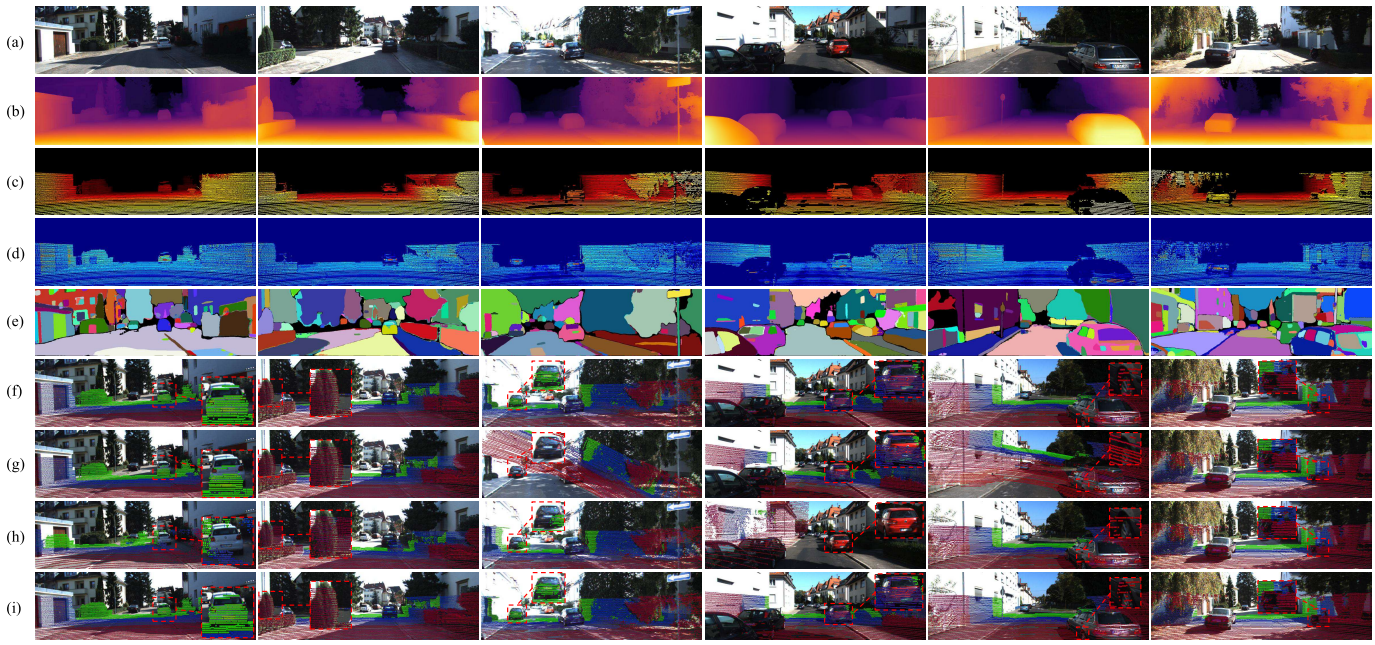

Fig. 7. Qualitative comparisons with SoTA target-free LCEC approaches on the KITTI odometry dataset: (a)-(e) RGB images, Depth images, LDP images, LIP images and image segmentation results; (f)-(i) experimental results achieved using ground truth, UMich, HKU-Mars and EdO-LCEC (ours), shown by merging LiDAR depth projections and RGB images, where significantly improved regions are shown with red dashed boxes.

TABLE I

QUANTITATIVE COMPARISONS WITH SOTA TARGET-FREE LCEC APPROACHES ON THE 00 SEQUENCE OF KITTI ODOMETRY. THE BEST RESULTS ARE SHOWN IN BOLD TYPE. †: THESE METHODS DID NOT RELEASE CODE, PREVENTING THE REPRODUCTION OF RESULTS FOR BOTH CAMERAS

Approach	Initial Range	Left Camera									Right Camera								
		Magnitude		Rotation Error (°)			Translation Error (m)				Magnitude		Rotation Error (°)			Translation Error (m)			
		e_r (°)	e_t (m)	Yaw	Pitch	Roll	X	Y	Z		e_r (°)	e_t (m)	Yaw	Pitch	Roll	X	Y	Z	
CalibRCNN† [26]	$\pm 10° / \pm 0.25m$	0.805	0.093	0.446	0.640	0.199	0.062	0.043	0.054		-	-	-	-	-	-	-	-	
CalibDNN† [27]	$\pm 10° / \pm 0.25m$	1.021	0.115	0.200	0.990	0.150	0.055	0.032	0.096		-	-	-	-	-	-	-	-	
RegNet† [44]	$\pm 20° / \pm 1.5m$	0.500	0.108	0.240	0.250	0.360	0.070	0.070	0.040		-	-	-	-	-	-	-	-	
LCCNet [25]	$\pm 10° / \pm 1.0m$	1.418	0.600	0.455	0.835	0.768	0.237	0.333	0.329		1.556	0.718	0.457	1.023	0.763	0.416	0.333	0.337	
RGGNet [29]	$\pm 20° / \pm 0.3m$	1.290	0.114	0.640	0.740	0.350	0.081	**0.028**	0.040		3.870	0.235	1.480	3.380	0.510	0.180	**0.056**	0.061	
CalibNet [28]	$\pm 10° / \pm 0.2m$	5.842	0.140	2.873	2.874	3.185	0.065	0.064	0.083		5.771	0.137	2.877	2.823	3.144	**0.063**	0.062	0.082	
Borer *et al.*† [24]	$\pm 1° / \pm 0.25m$	0.455	0.095	0.100	0.440	**0.060**	**0.037**	0.030	0.082		-	-	-	-	-	-	-	-	
CRLF [40]	-	0.629	4.116	**0.033**	0.464	0.416	3.648	1.473	0.550		0.633	4.606	**0.039**	0.458	0.424	4.055	1.636	0.644	
UMich [20]	-	4.161	0.319	0.113	3.111	2.138	0.286	0.068	0.086		4.285	0.329	0.108	3.277	2.088	0.290	0.085	0.090	
HKU-Mars [19]	-	33.84	6.354	19.89	18.71	19.32	3.353	3.228	2.419		32.89	4.913	18.99	15.77	17.00	2.917	2.564	1.646	
DVL [32]	-	122.1	5.129	48.64	87.29	98.15	2.832	2.920	1.881		120.5	4.357	49.60	87.99	96.72	2.086	2.517	1.816	
MIAS-LCEC [4]	-	5.385	1.013	1.574	4.029	4.338	0.724	0.373	0.343		7.655	1.342	1.910	5.666	6.154	0.843	0.730	0.358	
EdO-LCEC (Ours)	-	**0.295**	**0.078**	0.117	**0.176**	0.150	0.051	0.038	**0.032**		**0.336**	**0.118**	0.216	**0.168**	**0.121**	0.083	0.067	**0.032**	

TABLE II

COMPARISONS WITH SOTA LCEC APPROACHES ON KITTI ODOMETRY (01-09 SEQUENCES). THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Approach	01		02		03		04		05		06		07		08		09	
	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t
CRLF [40]	**0.623**	7.363	0.632	3.640	0.845	6.007	**0.601**	0.372	0.616	5.961	0.615	25.762	0.606	1.807	0.625	5.376	0.626	5.133
UMich [20]	2.196	**0.303**	3.733	0.329	3.201	0.316	2.086	0.348	3.526	0.356	2.914	0.353	3.928	0.368	3.722	0.367	3.117	0.363
HKU-Mars [19]	20.73	3.768	32.95	12.69	21.99	3.493	4.943	0.965	34.42	6.505	25.20	7.437	33.10	7.339	26.62	8.767	20.38	3.459
DVL [32]	112.0	2.514	120.6	4.285	124.7	4.711	113.5	4.871	123.9	4.286	128.9	5.408	124.7	5.279	126.2	4.461	116.7	3.931
MIAS-LCEC [4]	0.621	0.298	0.801	0.324	1.140	0.324	0.816	0.369	4.768	0.775	2.685	0.534	11.80	1.344	5.220	0.806	0.998	0.432
EdO-LCEC (Ours)	2.269	0.462	**0.561**	**0.137**	**0.737**	**0.137**	1.104	**0.339**	**0.280**	**0.093**	**0.485**	**0.124**	**0.188**	**0.076**	**0.352**	**0.115**	**0.386**	**0.120**

achieved by our method.Additionally, experimental results on the MIAS-LCEC-TF360 further prove our outstanding performance. From Table IV, it is evident that while the other approaches achieve poor performances, our method demonstrates excellent accuracy, indicating strong adaptability to more challenging scenarios, with narrow overlapping

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE III

COMPARISONS WITH SOTA TARGET-FREE LCEC APPROACHES ON MIAS-LCEC-TF70. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Approach	Residential Community		Urban Freeway		Building		Challenging Weather		Indoor		Challenging Illumination		All	
	e_r (°)	e_t (m)	e_r (°)	e_t (m)	e_r (°)	e_t (m)	e_r (°)	e_t (m)	e_r (°)	e_t (m)	e_r (°)	e_t (m)	e_r (°)	e_t (m)
CRLF [40]	1.594	0.464	1.582	0.140	1.499	20.17	1.646	2.055	1.886	30.05	1.876	19.05	1.683	11.13
UMich [20]	4.829	0.387	2.267	0.166	11.914	0.781	1.851	0.310	2.029	0.109	5.012	0.330	4.265	0.333
HKU-Mars [19]	2.695	1.208	2.399	1.956	1.814	0.706	2.578	1.086	2.527	0.246	14.996	3.386	3.941	1.261
DVL [32]	0.193	0.063	0.298	0.124	0.200	0.087	0.181	0.052	0.391	0.030	1.747	0.377	0.423	0.100
MIAS-LCEC [4]	0.190	0.050	**0.291**	0.111	0.198	0.072	**0.177**	0.046	0.363	**0.024**	0.749	0.118	0.298	0.061
EdO-LCEC (Ours)	**0.168**	**0.044**	0.293	**0.105**	**0.184**	**0.057**	0.183	**0.044**	**0.338**	0.027	**0.474**	**0.104**	**0.255**	**0.055**

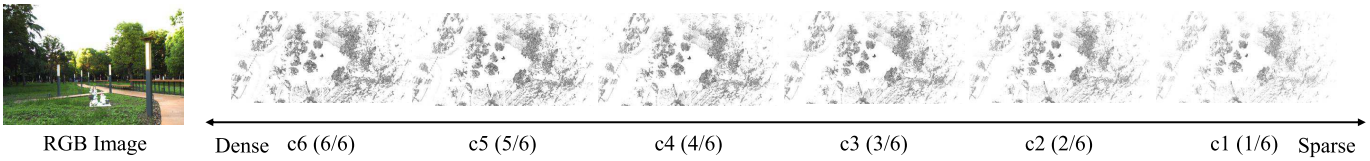

Fig. 8. We divided each point cloud into six equal parts and combined these segments to create point clouds with varying densities.

TABLE IV

QUANTITATIVE COMPARISONS OF OUR PROPOSED EDO-LCEC APPROACH WITH OTHER SOTA TARGET-FREE APPROACHES ON THE MIAS-LCEC-TF360. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Approach	Indoor		Outdoor	
	e_r (°)	e_t (m)	e_r (°)	e_t (m)
CRLF [40]	1.479	13.241	1.442	0.139
UMich [20]	1.510	0.221	6.522	0.269
HKU-Mars [19]	85.834	7.342	35.383	8.542
DVL [32]	39.474	0.933	65.571	1.605
MIAS-LCEC [4]	0.996	0.182	0.659	0.114
EdO-LCEC (Ours)	**0.720**	**0.106**	**0.349**	**0.109**

TABLE V

QUANTITATIVE COMPARISONS OF OUR PROPOSED EDO-LCEC APPROACH WITH OTHER SOTA TARGET-FREE APPROACHES ON THE NUSCENES DATASET. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Approach	Magnitude		Rotation Error (°)			Translation Error (m)		
	e_r (°)	e_t (m)	Yaw	Pitch	Roll	X	Y	Z
CRLF [40]	2.446	14.82	0.537	0.454	2.317	12.25	4.159	4.852
UMich [20]	4.340	0.390	0.524	2.118	3.185	0.297	**0.071**	0.200
HKU-Mars [19]	36.69	17.55	16.150	24.91	19.69	11.216	5.878	7.171
DVL [32]	108.9	3.371	42.50	71.71	70.88	1.672	1.874	1.380
MIAS-LCEC [4]	2.731	0.688	0.625	0.463	2.570	0.554	0.072	0.325
EdO-LCEC (Ours)	**1.090**	**0.261**	**0.414**	**0.356**	**0.842**	**0.202**	0.076	**0.100**

areas between LiDAR projections and camera images. This impressive performance can be attributed to our proposed DPCM, a powerful cross-modal feature matching algorithm. DPCM utilizes structural and textural consistency to jointly constrain correspondences matching on both spatial and textural pathways. This largely increases reliable correspondences compared to DVL and MIAS-LCEC, thereby providing a more reliable foundation for extrinsic parameter optimization.

3) Comparison With SoTA Approaches on nuScenes: Following the official split of nuScenes, we adopt the 150 test scenes in the v1.0-test set to construct the evaluation data. The experimental results, summarized in Table V, show that EdO-LCEC significantly outperforms prior methods. We attribute this improvement to the robust correspondences established by DPCM. Although existing approaches obtain dense correspondences from accumulated point clouds captured by solid-state LiDARs, the accuracy of these correspondences is limited and does not generalize well to unseen scenarios. The substantial modality gap between LiDAR point clouds and camera images further hinders prior methods from achieving high pose estimation accuracy. Moreover, when applied to sparse point clouds captured by mechanical LiDARs, these methods suffer a severe degradation in performance, largely due to the pronounced differences in feature density between LiDAR and camera data. In contrast, our approach leverages cross-modal segmented masks to bridge the modality gap. The corner points extracted from these masks inherently encode reliable semantic information, which is critical for effective correspondence matching. DPCM not only increases the number of correspondences but also improves their geometric distribution, making them better suited for accurate pose estimation.

C. Ablation Study and Analysis

To evaluate the algorithm's adaptability to incomplete and sparse point clouds, we further segmented the already limited field-of-view point clouds from the MIAS-LCEC-TF360 dataset. Specifically, as shown in Fig. 8, we divide each point cloud into six equal parts based on the recorded temporal sequence of the points. By progressively combining these segments, we create point clouds with varying densities, ranging from 1/6 (c1) to the full 6/6 (c6) density. The calibration results presented in Fig. 6 show that EdO-LCEC achieves the smallest mean e_r and e_t, along with the narrowest interquartile range, compared to other approaches across different point cloud densities. This demonstrates the stability and adaptability of

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE VI

ABLATION STUDY OF STRUCTURAL AND TEXTURAL CONSISTENCY ON THE 00 SEQUENCE OF KITTI ODOMETRY (MULTI-SCENE OPTIMIZATION IS NOT USED). THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Consistency		Left Camera		Right Camera	
Structural	Textural	e_r (°)	e_t (m)	e_r (°)	e_t (m)
		2.222	0.792	2.600	0.896
✓		1.227	0.304	1.533	0.411
	✓	2.416	0.827	2.633	0.984
✓	✓	**1.125**	**0.278**	**1.425**	**0.354**

EdO-LCEC under challenging conditions involving sparse or incomplete point clouds.

To validate the contribution of structural consistency and textural consistency in DPCM, we conducted an ablation study comparing their individual and combined effects. Multi-scene optimization is not used in this ablation study to better demonstrate the influence of the two consistencies. As shown in Table VI, when both are used, the two components significantly improve calibration accuracy. Specifically, structural consistency preserves local geometric features through mask-guided alignment, while textural consistency evaluates visual similarity around matched correspondences. Their combination enables DPCM to maintain robust matching performance even in challenging scenarios with sparse or occluded point clouds, which is a capability that existing methods (such as DVL and MIAS-LCEC) struggle to achieve. Additionally, MIAS-LCEC only compares the corner points within the matched segmented mask, while EdO-LCEC not only considers the corner points of the matched mask but also includes those of other unmatched masks. This allows EdO-LCEC to achieve global attention when searching for potential available matches when LiDAR frames are sparse.

Furthermore, we explore the contribution of the generalizable scene discriminator as well as the multi-view and multi-scene joint optimization. In this ablation study, the algorithm's performance is comprehensively evaluated on the 00 sequence of KITTI odometry and KITTI-360, both with and without multi-view optimization (including both textural and spatial perspective views) and multi-scene optimization. The results presented in Table VII and VIII demonstrates that lacking any of these components significantly degrades calibration performance. In particular, calibration errors increase substantially when only single-view calibration is applied, as it lacks the comprehensive constraints provided by multi-view inputs. Additionally, the joint optimization from multiple scenes significantly improved the calibration accuracy compared to that under only multi-view optimization. These results confirm the advantage of incorporating spatial and textural constraints from multiple views and scenes, validating the robustness and adaptability of our environment-driven calibration strategy.

D. Comparison of Correspondence Matching

We present a comparison of correspondence matching results in Fig. 9. The visualized results indicate that our

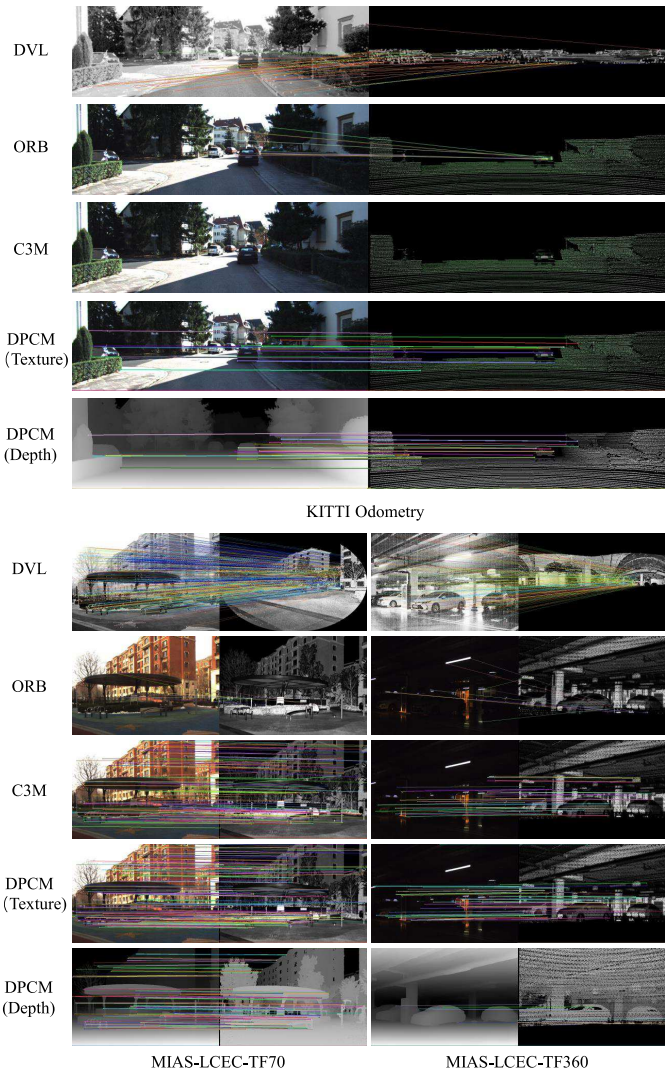

Fig. 9. Qualitative comparisons of correspondence matching.

method significantly outperforms other correspondence matching algorithms employed in the comparison calibration methods. Specifically, our approach yields substantially more correct correspondences than ORB, DVL (which employs SuperGlue for direct 3D-2D correspondence matching), and C3M (the cross-modal mask matching algorithm provided by MIAS-LCEC). This demonstrates the superior adaptability of DPCM, particularly in scenarios involving sparse point clouds with minimal geometric and textural information.

The ORB algorithm relies on the fast Harris corner detector for keypoint extraction and employs a binary descriptor for efficient matching based on Hamming distance. While ORB performs well in image-to-image matching, it struggles with cross-modal feature matching, especially in LiDAR-camera integration scenarios. Conversely, while DVL and C3M perform adequately on dense point clouds such as those in the MIAS-LCEC-TF70, they face significant challenges when applied to the sparse point clouds in the KITTI odometry and MIAS-LCEC-TF360. The narrow overlapping field of view between LiDAR and camera in KITTI odometry creates difficulties in identifying reliable correspondences, particularly

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

TABLE VII

ABLATION STUDY OF MULTI-VIEW MULTI-SCENE JOINT OPTIMIZATION ON THE 00 SEQUENCE OF KITTI ODOMETRY. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Components			Left Camera								Right Camera							
Multi-View		Multi	Magnitude		Rotation Error (°)			Translation Error (m)			Magnitude		Rotation Error (°)			Translation Error (m)		
Intensity	Depth	Scene	e_r (°)	e_t (m)	Yaw	Pitch	Roll	X	Y	Z	e_r (°)	e_t (m)	Yaw	Pitch	Roll	X	Y	Z
			1.625	0.457	0.820	0.669	0.899	0.247	0.232	0.211	1.620	0.472	0.915	0.691	0.807	0.257	0.254	0.197
✓			1.387	0.357	0.755	0.579	0.711	0.189	0.205	0.152	1.641	0.459	1.012	0.679	0.738	0.257	0.257	0.178
✓	✓		1.125	0.278	0.534	0.534	0.613	0.134	0.148	0.136	1.425	0.354	0.856	0.563	0.679	0.186	0.205	0.138
		✓	0.406	0.151	0.180	0.201	0.223	0.119	0.058	0.049	0.447	0.192	0.227	0.243	0.211	0.148	0.094	0.051
✓		✓	0.339	0.106	0.179	**0.167**	0.162	0.069	0.056	0.039	0.480	0.138	0.322	0.239	0.150	0.096	0.084	0.033
✓	✓	✓	**0.295**	**0.078**	**0.117**	0.176	**0.150**	**0.051**	**0.038**	**0.032**	**0.336**	**0.118**	**0.216**	**0.168**	**0.121**	**0.083**	**0.067**	**0.032**

TABLE VIII

ABLATION STUDY OF MULTI-VIEW MULTI-SCENE JOINT OPTIMIZATION ON THE 00 SEQUENCE OF KITTI-360. THE BEST RESULTS ARE SHOWN IN BOLD TYPE

Components			Camera 1		Camera 2	
Multi-View		Multi	Magnitude		Magnitude	
Intensity	Depth	Scene	e_r (°)	e_t (m)	e_r (°)	e_t (m)
			2.232	0.524	1.909	0.440
✓			1.832	0.402	1.882	0.415
✓	✓		1.496	0.323	1.506	0.311
		✓	0.914	0.197	0.686	0.134
✓		✓	0.755	0.135	0.612	0.085
✓	✓	✓	**0.605**	**0.079**	**0.498**	**0.061**

along the edges of the LiDAR field of view. This limitation significantly reduces the matching accuracy of DVL. Similarly, C3M, which restricts matching to corner points within aligned masks, struggles with sparse point clouds and limited sensor overlap.

Our proposed DPCM resolves these issues by incorporating both spatial and textural constraints. Unlike C3M, which treats mask instance matching as a strict constraint, DPCM uses it as a prior, enabling the generation of denser 3D-2D correspondences. This refinement significantly enhances the robustness of the algorithm in sparse environments and restricted fields of view, ensuring reliable calibration even under challenging conditions.

V. CONCLUSION

In this article, we explore extending a new definition called "environment-driven" for online LiDAR-camera extrinsic calibration. Unlike previous methods, our approach introduces environmental observation to maintain optimal performance across diverse and complex conditions. Specifically, we designed a scene discriminator that can automatically observe the calibration scene. This discriminator can guide cross-modal feature matching by evaluating the feature density through multi-modal features extracted by large vision models. By leveraging structural and textural consistency between LiDAR projections and camera images, our method achieves more reliable 3D-2D correspondence matching. Additionally, we modeled the calibration process as a multi-view and multi-scene joint optimization problem, achieving high-precision and robust extrinsic matrix estimation through multi-view optimization within individual scenes and joint optimization across multiple scenarios. Extensive experiments on real-world datasets demonstrate that our environment-driven calibration strategy achieves the state-of-the-art performance.

REFERENCES

[1] J. Jiao, F. Chen, H. Wei, J. Wu, and M. Liu, "LCE-calib: Automatic LiDAR-frame/event camera extrinsic calibration with a globally optimal solution," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 5, pp. 2988–2999, Oct. 2023.

[2] X. Chen et al., "Joint scene flow estimation and moving object segmentation on rotational LiDAR data," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 17733–17743, Nov. 2024.

[3] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Comput. Hum. Behav.*, vol. 122, Sep. 2021, Art. no. 106855.

[4] Z. Huang, Y. Zhang, Q. Chen, and R. Fan, "Online, target-free LiDAR-camera extrinsic calibration via cross-modal mask matching," *IEEE Trans. Intell. Vehicles*, vol. 10, no. 5, pp. 3531–3542, May 2025.

[5] G. P. Cruz et al., "EKF-LOAM: An adaptive fusion of LiDAR SLAM with wheel odometry and inertial data for confined spaces with few geometric features," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1458–1471, Jul. 2022.

[6] J. Lin and F. Zhang, "R3LIVE: A robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 10672–10678.

[7] Y. Yu, P. Yun, B. Xue, J. Jiao, R. Fan, and M. Liu, "Accurate and robust visual localization system in large-scale appearance-changing environments," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 6, pp. 5222–5232, Dec. 2022.

[8] F. Hui, Z. Zhou, and Y. Liu, "PL-LVI: A LiDAR-visual-inertial SLAM system integrating visual point-line features," *IEEE Trans. Autom. Sci. Eng.*, early access, 2025, doi: 10.1109/TASE.2025.3559668.

[9] Y. Tian, Y. Chang, F. Herrera Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic SLAM for multi-robot systems," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2022–2038, Aug. 2022.

[10] C. Shi, X. Chen, J. Xiao, B. Dai, and H. Lu, "Fast and accurate deep loop closing and relocalization for reliable LiDAR SLAM," *IEEE Trans. Robot.*, vol. 40, pp. 2620–2640, 2024.

[11] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 918–927.

[12] N. Wang et al., "SegNet4D: Efficient instance-aware 4D semantic segmentation for LiDAR point cloud," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 15339–15350, 2025.

[13] L. Luo et al., "BEVPlace++: Fast, robust, and lightweight LiDAR global localization for unmanned ground vehicles," *IEEE Trans. Robot.*, pp. 1–20, 2025.

[14] J. Cui, J. Niu, Z. Ouyang, Y. He, and D. Liu, "ACSC: Automatic calibration for non-repetitive scanning solid-state LiDAR and camera systems," 2020, *arXiv:2011.08516*.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

[15] G. Yan, F. He, C. Shi, P. Wei, X. Cai, and Y. Li, "Joint camera intrinsic and LiDAR-camera extrinsic calibration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11446–11452.

[16] J. Beltrán, C. Guindel, A. de la Escalera, and F. García, "Automatic extrinsic calibration method for LiDAR and camera sensor setups," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17677–17689, Oct. 2022.

[17] Y. Xie et al., "A4LiDARTag: Depth-based fiducial marker for extrinsic calibration of solid-state LiDAR and camera," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6487–6494, Jul. 2022.

[18] H. Huang, M. Zhang, L. Li, J. Hu, and H. Wang, "GTSCalib: Generalized target segmentation for target-based extrinsic calibration of non-repetitive scanning LiDAR and camera," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 3648–3660, 2025.

[19] C. Yuan, X. Liu, X. Hong, and F. Zhang, "Pixel-level extrinsic self calibration of high resolution LiDAR and camera in targetless environments," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7517–7524, Oct. 2021.

[20] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic extrinsic calibration of vision and LiDAR by maximizing mutual information," *J. Field Robot.*, vol. 32, no. 5, pp. 696–722, Aug. 2015.

[21] Y. Wang et al., "Automatic registration of point cloud and panoramic images in urban scenes based on pole matching," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, Dec. 2022, Art. no. 103083.

[22] Y. Han, Y. Liu, D. Paz, and H. Christensen, "Auto-calibration method using stop signs for urban autonomous driving applications," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13179–13185.

[23] Y. Liao et al., "SE-calib: Semantic edge-based LiDAR–camera boresight online calibration in urban scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 1000513.

[24] J. Borer, J. Tschirner, F. Ölsner, and S. Milz, "From chaos to calibration: A geometric mutual information approach to target-free camera LiDAR extrinsic calibration," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8394–8403.

[25] X. Lv, B. Wang, Z. Dou, D. Ye, and S. Wang, "LCCNet: LiDAR and camera self-calibration using cost volume network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2894–2901.

[26] J. Shi et al., "CalibRCNN: Calibrating camera and LiDAR by recurrent convolutional neural network and geometric constraints," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10197–10202.

[27] G. Zhao, J. Hu, S. You, and C. C. J. Kuo, "CalibDNN: Multimodal sensor calibration for perception using deep neural networks," in *Proc. Signal Process., Sensor/Inf. Fusion, Target Recognit. XXX*, Apr. 2021, pp. 324–335.

[28] G. Iyer, R. K. Ram, J. K. Murthy, and K. M. Krishna, "CalibNet: Geometrically supervised extrinsic calibration using 3D spatial transformer networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1110–1117.

[29] K. Yuan, Z. Guo, and Z. J. Wang, "RGGNet: Tolerance aware LiDAR-camera online calibration with geometric deep learning and generative model," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6956–6963, Oct. 2020.

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

[31] C. Zhang et al., "Faster segment anything: Towards lightweight SAM for mobile applications," 2023, *arXiv:2306.14289*.

[32] K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "General, single-shot, target-less, and automatic LiDAR-camera extrinsic calibration toolbox," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11301–11307.

[33] N. Ou, H. Cai, J. Yang, and J. Wang, "Targetless extrinsic calibration of camera and low-resolution 3-D LiDAR," *IEEE Sensors J.*, vol. 23, no. 10, pp. 10889–10899, May 2023.

[34] F. Lv and K. Ren, "Automatic registration of airborne LiDAR point cloud data and optical imagery depth map based on line and points features," *Infr. Phys. Technol.*, vol. 71, pp. 457–463, Jul. 2015.

[35] J. Castorena, U. S. Kamilov, and P. T. Boufounos, "Autocalibration of LiDAR and optical cameras via edge alignment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2862–2866.

[36] S. Tang et al., "Robust calibration of vehicle solid-state LiDAR-camera perception system using line-weighted correspondences in natural environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 4489–4502, May 2024.

[37] D. Zhang, L. Ma, Z. Gong, W. Tan, J. Zelek, and J. Li, "An overlap-free calibration method for LiDAR-camera platforms based on environmental perception," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–7, 2023.

[38] J. Yin, F. Yan, Y. Liu, and Y. Zhuang, "Automatic and targetless LiDAR–camera extrinsic calibration using edge alignment," *IEEE Sensors J.*, vol. 23, no. 17, pp. 19871–19880, Sep. 2023.

[39] N. Ou, H. Cai, and J. Wang, "Targetless LiDAR-camera calibration via cross-modality structure consistency," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 2636–2648, Jan. 2024.

[40] T. Ma, Z. Liu, G. Yan, and Y. Li, "CRLF: Automatic calibration and refinement based on line feature for LiDAR and camera in road scenes," 2021, *arXiv:2103.04558*.

[41] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4937–4946.

[42] L. Yang et al., "Depth anything V2," 2024, *arXiv:2406.09414*.

[43] A. M. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4015–4026.

[44] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "RegNet: Multimodal sensor registration using deep neural networks," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1803–1810.

[45] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023.

[46] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

Zhiwei Huang received the B.E. degree in automation from Tongji University, Shanghai, China, in 2024, where he is currently pursuing the M.Sc. degree with the MIAS Group, College of Electronic and Information Engineering, supervised by Prof. Rui Fan. His research interests include computer vision and robotics.

Jiaqi Li is currently pursuing the B.E. degree in automation with Tongji University, Shanghai, China. His academic focus is on computer vision and autonomous systems.

Hongbo Zhao received the B.S. degree in mathematics from Tongji University, Shanghai, China, in 2024, where he is currently pursuing the Ph.D. degree with the MIAS Group, Shanghai Research Institute for Intelligent Autonomous Systems, supervised by Prof. Rui Fan. His research interests include pose estimation and 3D reconstruction in computer vision.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2026, Vienna, Austria.

Xiao Ma received the M.S. degree in micro-electro-mechanical systems (MEMS) from the School of Electronics and Computer Science, University of Southampton, U.K., in 2018. Since 2019, he has been with Beijing Institute of Aerospace Control Devices, where he is currently a Microsystem Designer. His research interests include MEMS inertial instruments, wafer-level packaging (WLP), and system-in-package (SiP) technologies.

Ping Zhong (Member, IEEE) received the B.S. degree from the National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree in communication engineering from Xiamen University, China, in 2011. She is currently an Associate Professor with the Department of Computer Science and Technology, Central South University. Her research interests include autonomous systems, machine learning, and the Internet of Things.

Xiao-Hu Zhou (Member, IEEE) received the B.E. degree in automation from Central South University, Changsha, China, in 2014, and the Ph.D. degree in control theory and control engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently a Professor with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include surgical robotics, skill learning, and deep learning.

Wei Ye (Member, IEEE) received the Ph.D. degree in computer science from the Institut für Informatik, Ludwig-Maximilians-Universität München, Munich, Germany, in 2018. He is currently a tenure-track Professor with the College of Electronic and Information Engineering, Tongji University, Shanghai, China, the Frontiers Science Center for Intelligent Autonomous Systems (Ministry of Education), Shanghai, and Shanghai Innovation Institute, Shanghai.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in automation from Harbin Institute of Technology in 2015 and the Ph.D. degree in electrical and electronic engineering from the University of Bristol in 2018. He worked as a Research Associate at The Hong Kong University of Science and Technology from 2018 to 2020 and a Post-Doctoral Scholar-Employee at the University of California San Diego from 2020 to 2021. He began his faculty career as a Full Research Professor with the College of Electronic and Information Engineering, Tongji University, in 2021. He was promoted to a Full Professor and attained Tenure with the College of Electronic and Information Engineering and Shanghai Research Institute for Intelligent Autonomous Systems in 2022 and 2024, respectively. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis. He served as a Senior Program Committee Member for AAAI'23/24/25/26. He organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21/25, and ECCV'22. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide between 2022 and 2025, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, acknowledged as one of Xiaomi Young Talents in 2023, and awarded Shanghai Science and Technology 35 Under 35 Honor in 2024 as its youngest recipient. He served as the Area Chair for ICIP'24. He served as an Associate Editor for ICRA'23/25 and IROS'23/24.