2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MIND-Calib: Multi-View, Intensity and Depth-driven Dense 2D–3D Alignment for Single-Frame LiDAR–Camera Extrinsic Calibration

Shezhong Liu[1]*† and Zibin Chen[12]*

***Abstract*— Extrinsic calibration between LiDAR and camera is a crucial step in multi-sensor fusion, where targetless approaches have attracted increasing attention for their flexibility and reusability. However, existing methods still suffer from three major limitations: time-consuming data preparation, lack of robustness under sparse single-frame input, and limited generalization across diverse LiDAR architectures. We propose MIND-Calib, a truly single-frame, targetless calibration framework. The method generates depth and intensity images through virtual multi-view projection, and performs image-domain completion and back-projection to densify the point cloud and construct sub-pixel 2D–3D correspondences. High-precision extrinsics are then estimated via dual-channel cross-modal matching that leverages both depth and intensity modalities. Experiments on three representative LiDAR types (MEMS-based, solid-state, and mechanical spinning) as well as on public datasets demonstrate an average accuracy of 2.85 cm (with respect to an average scene depth of 40 meters) in translation and 0.20° in rotation. More importantly, MIND-Calib not only achieves true single-frame calibration without any additional preparation, but also maintains stable accuracy under sparse inputs and exhibits strong generalization and robustness across devices and challenging environments.**

I. Introduction

Accurate environmental perception is essential for autonomous driving and robotics [1], [2]. LiDAR–camera calibration directly affects multimodal fusion quality [3]. Traditional target-based calibration methods achieve high accuracy but involve cumbersome procedures and are less convenient for addressing long-term drift or sensor failures [4], motivating the rise of targetless approaches.

Most existing targetless approaches rely on natural scene features [5], [6] or deep learning models [7], [8] to achieve automatic calibration. Koide *et al.* proposed the LiDAR–Camera Extrinsic Calibration Toolbox [9], which estimates extrinsics through point cloud projection and feature matching with images. However, this method strongly depends on point cloud density, leading to substantial accuracy degradation or outright failure under sparse data.

MIAS-LCEC [10] employs MobileSAM [11] to extract segmentation masks from both RGB and LiDAR projection images and uses a two-stage cross-modal mask matching (C3M) strategy for point-to-point alignment, showing potential for targetless and online calibration. Nonetheless, segmentation errors become significant in sparse projections, corner extraction is unstable, and the method relies on regular-shaped objects.

Fig. 1. Cross-modal feature correspondences between LiDAR point cloud and camera image, with zoomed-in regions highlighting accurate matches.

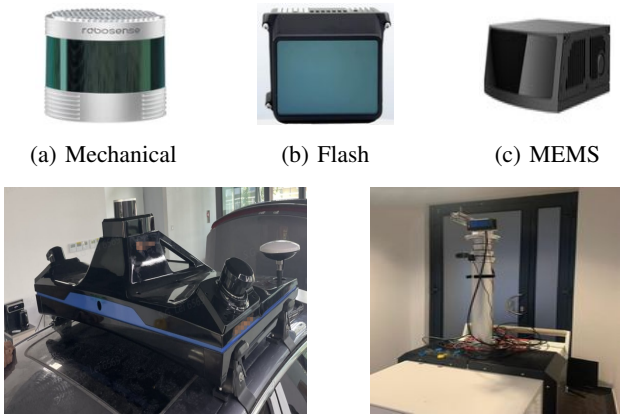
Overall, these methods remain challenged by data sparsity, scene complexity, and strong prior assumptions. In an effort to overcome these challenges, recent studies have turned to insights from other fields, investigating more resilient and broadly applicable approaches to cross-modal alignment.

Recent advances in Cross-Modal Image matching [12], [13] provide new inspiration. Such techniques have demonstrated robust registration between RGB, infrared, event camera, and remote sensing images even under severe texture loss or structural discrepancy. This progress suggests a novel perspective for LiDAR–camera calibration: matching in the image domain, which is more consistent with human intuition for spatial and geometric alignment.

Inspired by this trend, we propose a calibration method based on virtual multi-view and dual-channel cross-modal matching using only a single snapshot (see Fig. 1). Specifically, we construct multiple virtual camera viewpoints on the same LiDAR frame to generate depth and intensity images, which are then matched to the RGB image using state-of-the-art cross-modal image matching models (MINIMA) [12], enabling robust correspondence establishment across modalities. This multi-view design substantially improves spatial coverage and distribution uniformity of correspondences, strengthening the geometric constraints and mitigating local degeneration.

Furthermore, we introduce a geometry completion mechanism [14], [15] to address the limitations of sparse point clouds. We achieve geometric densification of the point cloud by leveraging depth and intensity information in the image domain. The depth map provides global scale constraints, while the intensity map contributes rich texture and boundary details, laying a solid foundation for precise cross-modal

*Equal contribution.
[1]Technical University of Berlin, Germany. {shezhong.liu.1, zibin.chen}@campus.tu-berlin.de
[2]LiangDao GmbH, Berlin, Germany, now invested by Agile Robots.
† Corresponding author.



(a) Mechanical (b) Flash (c) MEMS

(d) Vehicle platform (e) Road-side platform

Fig. 2. Various LiDAR types and mobile platforms used to validate our method.

alignment. The completed dense point cloud not only improves spatial continuity but also supports sub-pixel-level 2D–3D back-projection, significantly boosting matching accuracy and solution stability (see Fig. 1, right).

We validated our method on different LiDAR sensor types and multiple actual system platforms, as shown in Fig. 2, including a vehicle-mounted system with multi-camera/LiDAR integration (see Fig. 2d) and a roadside system with sensors on a 4 m telescopic pole (see Fig. 2e). The results demonstrate that our approach maintains consistently high matching accuracy and excellent stability across all configurations.

In this work, we operate on quasi-static LiDAR scans, where motion-induced rolling distortion is negligible. Such static snapshots provide high-quality geometric measurements and avoid introducing motion distortion as a systematic error into the calibration optimization. Handling scenarios with significant motion distortion would require an additional deskewing module (e.g., IMU- or odometry-assisted correction), which is beyond the scope of this paper.

The main contributions of this paper are summarized as follows:

- **Single-frame, zero-acquisition calibration framework (Instant Calibration):** Our method achieves high-precision geometric alignment without relying on trajectories, device motion, initialization, or additional data preparation. Calibration can be instantly performed upon system startup as long as the LiDAR and camera share an overlapping field of view.
- **Virtual Multi-View Dual-Channel Constraints:** By constructing globally distributed observations from multiple virtual viewpoints and leveraging the complementary properties of depth and intensity images, we employ state-of-the-art cross-modal image matching models to achieve stable and structurally consistent matching in the image domain. This significantly improves robustness under sparse or complex conditions, yielding abundant, uniformly distributed, and reliable 2D–2D correspondences.
- **Non-learning geometric densification with depth–intensity:** By completing the projected depth and intensity images in the image domain and back-projecting them into 3D space, we generate a densified and structurally consistent point cloud. This process enables sub-pixel accurate 2D–3D correspondences by more precisely locating the 3D points behind the 2D projections, thereby substantially reducing geometric error.

II. Related Work

Extrinsic calibration between LiDAR and camera is a fundamental task in multi-sensor fusion systems, aiming to estimate the rigid-body transformation between the two modalities. Existing methods can be broadly categorized into target-based and target-less approaches.

A. Target-based Methods

Target-based calibration methods rely on manually placed calibration objects such as checkerboards, dot patterns to construct strong geometric correspondences [16], [17] (e.g., point-to-point or point-to-plane), achieving sub-pixel accuracy in controlled environments. However, these methods require manual intervention, complex data acquisition procedures, and specific scene constraints, making them difficult to automate in practical systems.

B. Target-less Methods

To improve flexibility and deployment efficiency, recent research has shifted toward automated calibration methods without artificial targets. These approaches can be categorized into the following three types:

a) ***Information-theoretic Methods****:* These methods estimate extrinsic parameters by maximizing statistical consistency between LiDAR and image modalities, such as reflectance–intensity, surface normal–gradient, or depth edge–image edge [18], [19]. For example, Luo et al. [20] proposed a zero-training LiDAR–camera calibration method (Calib-Anything), which generates category-free segmentation masks via SAM [21] and aligns the distribution of LiDAR intensity and normals with the image masks for self-supervised optimization. This approach avoids explicit feature engineering and shows strong generalization, but it remains sensitive to initialization and point cloud sparsity, and may degrade in complex or reflective environments.

b) ***Feature-based Methods****:* These approaches extract geometric features such as edges, corners, planes, or contours from both LiDAR and images [22], then establish correspondences for optimization via PnP (Perspective-n-Point) and iterative closest point, or nonlinear solvers. For instance, Yuan et al. [23] proposed a pixel-level extrinsic self-calibration method that aligns image and LiDAR edges via directional gradient consistency, achieving high accuracy with dense data but degrading under sparse or textureless conditions.

c) ***Deep Learning Methods****:* Deep learning-based methods learn cross-modal mappings using neural networks, either for direct extrinsic regression or as part of a hybrid pipeline. CMRNext [24] learns pixel-wise shift fields to

align LiDAR and image domains, offering high accuracy but requiring large-scale annotated data. Due to differences in LiDAR types, scan patterns, and intensity encoding, such models often lack cross-device generalization and require frequent retraining. Based on practical experience, the data structures and representations of different LiDARs vary significantly, which further limits model transferability.

To reduce deployment cost, MDPCalib [25] combines 2D–3D correspondences from CMRNext with motion estimation from visual or LiDAR odometry to perform geometric optimization without retraining. This method integrates automation and interpretability, but its final performance remains influenced by the quality of pretrained networks and matching robustness in complex environments.

III. Method

A. Overview

The overall workflow is illustrated in Fig. 3. The LiDAR point cloud is first projected into multiple virtual viewpoints, generating depth (MV-DEP) and intensity (MV-LIP) images as complementary modalities. These projections are then matched with the RGB image to establish cross-modal 2D–2D correspondences. Matched pixels are back-projected to recover 3D points, where a geometry-based densification step alleviates sparsity and improves reliability. Finally, extrinsic parameters are estimated through a two-stage PnP procedure with RANSAC outlier rejection and non-linear refinement.

B. Multi-view Generation

Several virtual cameras are constructed around the actual camera installation, with their optical centers uniformly sampled and orientations adjusted to cover the entire RGB image domain. These complementary projections yield denser and more uniformly distributed correspondences. Formally, let $\mathcal{C} = \{C_k\}_{k=1}^{K}$ denote the set of K virtual cameras, where each $C_k = (\mathbf{R}_k, \mathbf{t}_k)$ is defined with its optical center sampled around the real camera position and its orientation adjusted to ensure full coverage of the RGB image domain:

$$C_k = (\mathbf{R}_k, \mathbf{t}_k), \quad \mathbf{t}_k \sim \mathcal{U}(\mathcal{N}_{\text{cam}}). \tag{1}$$

where $\mathcal{U}(\cdot)$ denotes uniform sampling and $\mathcal{N}_{\text{cam}}$ is a local neighborhood around the actual camera installation. For each LiDAR point $\mathbf{p}_i \in \mathbb{R}^3$, its projection onto the k-th virtual camera image plane is obtained by

$$\tilde{\mathbf{u}}_{i,k} = \mathbf{K}(\mathbf{R}_k \mathbf{p}_i + \mathbf{t}_k), \tag{2}$$

where $\mathbf{K}$ is the intrinsic matrix of the virtual camera. The union of all projections across K views forms a densified 2D–3D correspondence set:

$$\mathcal{M} = \bigcup_{k=1}^{K} \{(\tilde{\mathbf{u}}_{i,k}, \mathbf{p}_i)\}. \tag{3}$$

As shown in Fig. 4, this multi-view construction balances spatial coverage and emphasizes distinctive local features, thereby providing stronger geometric constraints for calibration.

C. Dual-channel Projection

At this stage, we generate not only the MV-LIP but also the MV-DEP representations. As illustrated in Fig. 5, depth provides structural geometry, while intensity captures local detail, offering complementary information for calibration. Their integration not only reinforces geometric–photometric consistency but also significantly improves the robustness and accuracy of cross-modal feature matching, resulting in a more complete and reliable scene representation.

D. Cross-modal Feature Matching

To establish robust correspondences between the LiDAR projection and the RGB image, we adopt the Modality-Invariant Image Matching (MINIMA) framework [12], a state-of-the-art cross-modal matching algorithm. Unlike traditional feature-based methods that rely on handcrafted designs, MINIMA leverages a cross-modal data generation engine to construct a large-scale training dataset, covering six additional modalities (infrared, depth, event, normal, sketch, and paint). This design enables it to produce more accurate correspondences across diverse sensor inputs. Since our LiDAR intensity projection resembles event data and we also employ depth images, we directly utilize the released code and pretrained weights from this work. These reliable matches serve as the initial correspondences for our calibration pipeline. Specifically, we perform cross-modal matching independently between the MV-LIP image and the RGB image, and between the MV-DEP image and the RGB image, using the same MINIMA backbone. This produces two sets of 2D–2D correspondences, which are subsequently merged into a unified candidate set.

E. Back-projection and Point Cloud Densification

2D–2D correspondences must be lifted into 2D–3D. However, due to the sparsity of LiDAR point clouds, nearest-neighbor approximation often introduces large errors at object boundaries.

To mitigate this problem, we propose a geometry-consistent point cloud densification method inspired by classical depth completion approaches [14]. Unlike learning-based methods [26] that rely on modality-specific training data, classical interpolation operates directly on raw LiDAR measurements and thus maintains strong generalization across sensors with different resolutions and scanning patterns [27]. However, conventional interpolation methods often lead to over-smoothing and blurred boundaries. To address this, we propose the following joint-weight-based approach. In the virtual camera domain, we complete missing depth pixels as follows: for each unknown pixel, valid depth samples are first collected from their local neighborhood, joint weights are then computed based on intensity similarity (assigning larger weights to pixels with closer grayscale values) and depth proximity (favoring samples closer to the nearest reference depth). The missing depth value is finally obtained by weighted averaging. After completion, the depth map is back-projected into 3D space and further refined by

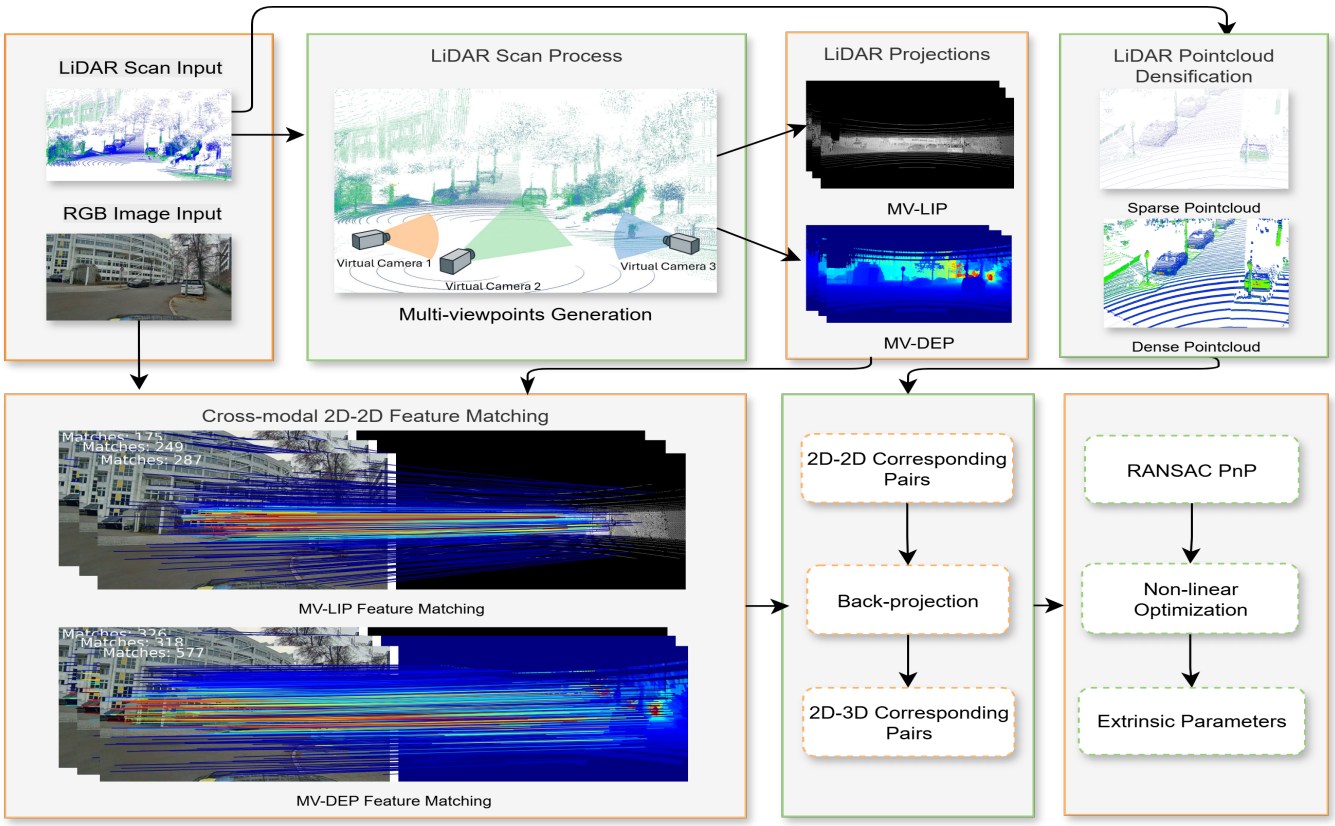
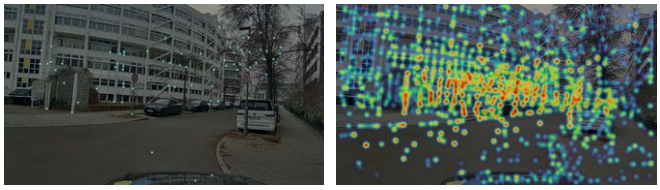
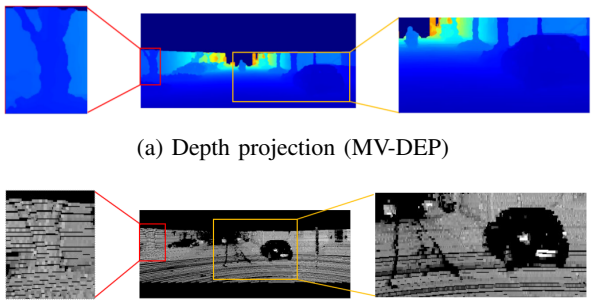


Fig. 3. Our pipeline calibrates from a single RGB–LiDAR pair. LiDAR scans are projected into multiple virtual views to generate intensity and depth maps, which are matched with the RGB image. After point cloud densification, reliable 2D–3D correspondences are recovered and extrinsics are estimated via PnP.



(a) Single viewpoint (b) 18 viewpoints

Fig. 4. Heatmap of matched point distributions: (a) single viewpoint and (b) 18 viewpoints, where multiple views produce denser and more uniform correspondences.



(a) Depth projection (MV-DEP)

(b) Intensity projection (MV-LIP)

Fig. 5. Depth and intensity projections provide complementary cues: depth captures reliable geometry (e.g., trees), while intensity emphasizes appearance details (e.g., lane markings and vehicles) that are poorly represented in depth.

ray-ghost [15] suppression and statistical filtering, yielding a dense and clean point cloud.

Formally, the completion weights are defined by intensity and depth consistency:

$$w_I = \exp\left(-\frac{(I - I_n)^2}{2\sigma_I^2}\right), \tag{4}$$

where pixels with similar intensity receive higher weights, encouraging interpolation only among appearance-consistent regions.

$$w_d = \exp\left(-\frac{(d_n - d_{\text{ref}})^2}{2\sigma_d^2}\right), \tag{5}$$

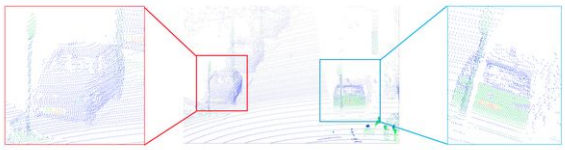
where neighbors closer to the reference depth are preferred, reducing errors across depth discontinuities.

The final completed depth is obtained via joint Gaussian weighting:

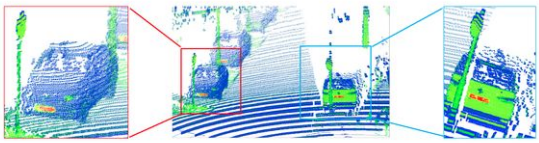
$$\hat{d}(u,v) = \frac{\sum_n w_n\, d_n}{\sum_n w_n}, \quad w_n = w_{I,n} \cdot w_{d,n}. \tag{6}$$

This ensures that only neighbors consistent in both intensity and depth contribute significantly, producing smooth yet boundary-preserving depth estimates.

Here, I and I_n denote the intensity values of the target pixel (u, v) and its neighboring pixel n in the virtual camera



(a) Before densification



(b) After densification

Fig. 6. Comparison of LiDAR point clouds before and after densification. The densified result shows improved depth continuity and sharper structural details, with intensity information also being better preserved.

view, respectively. d_n is the depth of the neighboring pixel, while d_{ref} denotes the nearest available reference depth at (u, v). The parameters σ_I and σ_d control the sensitivity to intensity and depth differences, respectively.

Once the depth is calculated, it can be back-projected through the virtual camera model to obtain the corresponding 3D point in the LiDAR coordinate system. This step produces a densified point cloud with sharper object boundaries, as illustrated in Fig. 6. The points are colored by intensity, demonstrating that the densification not only yields a higher-quality point cloud but also faithfully retains both geometric detail and intensity information—for instance, the high-reflectivity red license plate is clearly preserved.

F. PnP-based Extrinsic Estimation

With reliable 2D–3D correspondences, extrinsic calibration is solved as a standard PnP problem. We apply a RANSAC-based solver to discard outliers and then refine the solution with non-linear optimization, obtaining the final LiDAR–camera transformation.

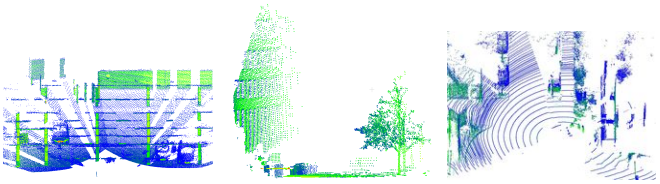
IV. Experimental Evaluation

A. Datasets and Ground Truth

We evaluate our method on two types of datasets: self-collected datasets and the public KITTI dataset [28].

Self-collected datasets: To comprehensively evaluate the robustness of the proposed method, we collected data (static single-frame) using three representative types of LiDAR sensors:

- *Mechanical spinning LiDAR*: Robosense Ruby128 (vehicle platform; outdoor and indoor) and Ouster OS1-64 (road-side platform; outdoor);
- *Solid-state (flash) LiDAR*: LD Satellite GenII Lite (road-side platform; outdoor);
- *MEMS-based LiDAR* (Micro-Electro-Mechanical Systems): Robosense M1 (vehicle platform; outdoor).



(a) (b) (c)

Fig. 7. Representative point cloud examples of three LiDAR types illustrating distinct scanning patterns: (a) MEMS-based, (b) Flash solid-state, (c) Mechanical spinning.

These three LiDARs exhibit distinct point cloud patterns, as shown in Fig. 7. They differ in field of view and point cloud resolution.

By collecting data from these three types of LiDAR sensors across diverse scenarios, we established a comprehensive self-collected dataset that provides a solid foundation for evaluating the stability and generalization capability of the proposed calibration framework.

KITTI dataset: We further validate our method on the KITTI odometry dataset. Following the official protocol, we use the provided LiDAR–camera extrinsics as ground truth.

Ground truth generation: For our self-collected datasets, the ground-truth extrinsics are obtained using a calibration target. The target provides 3D corner points in the LiDAR frame and corresponding 2D corner points in the image, enabling precise calibration.

All experiments are implemented in Python and executed on a workstation equipped with an Intel i7 CPU and an NVIDIA RTX 3070 GPU.

B. Quantitative Results

1) Baseline Methods: To provide a fair and thorough evaluation, we compare our approach against four representative baseline algorithms: HKU-Mars-Lab [23], Calib-Anything [20], MIAS-LCEC [10], and Ts et al. [29]. Their main characteristics are summarized in Table I.

To further validate the effectiveness of our method, we conducted quantitative comparisons on both self-collected and public datasets, covering different LiDAR sensor types.

2) Comparison with baseline algorithms on Ruby128 and KITTI datasets: Table II summarizes results on Ruby128 (indoor/outdoor) and KITTI. Our method consistently achieves the best performance. On the KITTI dataset, it attains a translation error of 1.53 cm and a rotation error of 0.184°, the lowest among all baselines and demonstrating strong generalization. Across all tested datasets, our calibration succeeds reliably and maintains stable.

3) Adaptability across LiDAR types: To further evaluate generalization, we tested our framework on three representative LiDAR types: MEMS (Robosense M1), solid-state (Satellite GenII Lite), and mechanical spinning (Ouster OS1-64). As shown in Table III, our method consistently achieves centimeter-level translation accuracy (below 4 cm) and rotation errors of around 0.3°, demonstrating robust adaptability across diverse LiDAR architectures and environments.

TABLE I

BASELINE METHODS USED FOR COMPARISON.

Method	Overview
HKU-Mars-Lab [23]	Edge-based matching using high-resolution LiDAR (e.g., Livox); achieves calibration in structured scenes.
Calib-Anything [20]	Uses SAM to segment images; compares LiDAR projection statistics (reflectivity, normals) within masks.
MIAS-LCEC [10]	Uses MobileSAM to segment RGB and LIP images; performs mask matching with a coarse-to-fine C3M algorithm.
Ts et al. [29]	Employs semantic segmentation of both images and LiDAR point clouds; performs lidar-to-camera registration to automate extrinsic parameter estimation.

All baseline algorithms exhibit poor convergence and large error margins when applied to LiDARs with different scanning architectures. Although Calib-Anything [20] shows relatively better robustness, its results still suffer from evident global misalignment in the LiDAR-to-image projections. As such misalignments are unacceptable for downstream tasks, these results are excluded from Table III.

TABLE II

QUANTITATIVE COMPARISON OF OUR METHOD AND BASELINE ALGORITHMS ON SELF-COLLECTED (ROBOSENSE RUBY128) AND PUBLIC (KITTI) DATASETS. FOR REFERENCE, THE AVERAGE SCENE DEPTH IS APPROXIMATELY 35 M IN THE INDOOR SETTING AND 50 M IN THE OUTDOOR SETTING.

Dataset	Scenario	Method	Trans. error (cm)	Rot. error(°)
Ruby128	Indoor	HKU-Mars-Lab [23]	29.3	1.221
		Calib-Anything [20]	10.2	0.370
		MIAS-LCEC [10]	7.1	0.389
		Ours	**4.0 ± 0.3**	**0.298 ± 0.03**
	Outdoor	HKU-Mars-Lab [23]	170.2	17.282
		Calib-Anything [20]	29.0	0.713
		MIAS-LCEC [10]	51.8	1.528
		Ours	**2.7 ± 0.5**	**0.206 ± 0.05**
KITTI	Outdoor	Ts et al. [29]	20.2	0.340
		HKU-Mars-Lab [23]	-	-
		Calib-Anything [20]	9.80	0.302
		MIAS-LCEC [10]	-	-
		Ours	**1.53 ± 1**	**0.184 ± 0.03**

TABLE III

MATCHING PERFORMANCE OF OUR METHOD ACROSS DIFFERENT LIDAR TYPES.

LiDAR Type	Trans. error (cm)	Rot. error (°)
M1 (MEMS)	**2.52 ± 0.3**	**0.12 ± 0.05**
Satellite (Solid-state)	**2.2 ± 0.5**	**0.183 ± 0.05**
OS1-64 (Spinning)	**3.83 ± 0.4**	**0.31 ± 0.04**

C. Qualitative Results

Fig. 1 illustrates the valid 2D–3D correspondence pairs used in the final extrinsic calibration, showing consistent alignment across views.

To demonstrate the robustness and generalizability of our calibration method across different LiDAR sensor types, we present qualitative results on four outdoor scenes, each captured using a distinct LiDAR configuration: a spinning mechanical LiDAR, a solid-state LiDAR, a MEMS-based LiDAR, and the Velodyne HDL-64E from the KITTI dataset. These sensors vary significantly in terms of scanning mechanisms, angular resolution, and point density. Nevertheless, our method produces consistent and accurate geometric alignment between LiDAR point clouds and image structures. As shown in Fig. 8, the calibrated projections align well with scene edges and semantic boundaries across all sensor types.

D. Ablation Studies

To evaluate the effectiveness of each component in improving the extrinsic calibration performance of **MIND-Calib**, we conducted a systematic ablation study on the KITTI dataset. The study incrementally incorporates three design aspects:

- **Viewpoint construction:** single-view vs. multi-view;
- **Geometric densification:** original sparse point cloud vs. densified point cloud;
- **Channel configuration:** intensity-only (LIP), depth-only (DEP), and the proposed dual-channel (LIP+DEP).

1) Evaluation Metrics: To more clearly assess the effectiveness of the proposed ablation factors, we introduce two auxiliary metrics: Retention Rate and Completeness.

Retention Rate: Retention Rate measures how many of the initial tentative correspondences remain after PnP-RANSAC filtering:

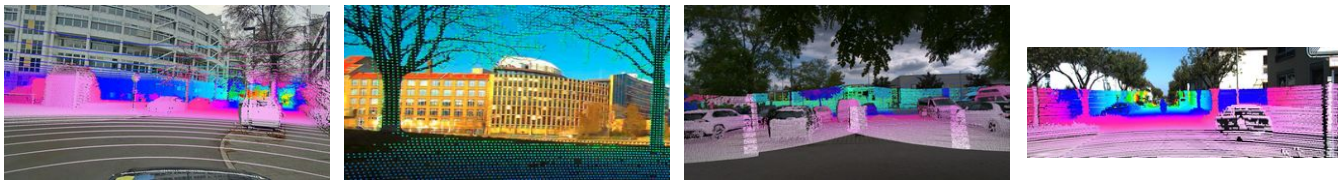
$$\text{Retention Rate} = \frac{|\mathcal{M}_{\text{inlier}}|}{|\mathcal{M}_{\text{tentative}}|}, \tag{7}$$

where $\mathcal{M}_{\text{tentative}}$ denotes all tentative 2D–3D correspondences before RANSAC, and $\mathcal{M}_{\text{inlier}}$ denotes the subset preserved as inliers. A higher retention rate indicates that more correspondences are preserved for the subsequent non-linear PnP optimization, thereby providing stronger geometric constraints and leading to more accurate extrinsic calibration.

Completeness: Completeness evaluates the spatial coverage of valid correspondences. The image is divided into a fixed 32×16 grid, and completeness is defined as:

$$\text{Completeness} = \frac{N_{\text{valid}}}{N_{\text{grid}}}, \tag{8}$$

where N_{valid} is the number of grid cells containing at least one valid correspondence, and N_{grid} is the total number of grid cells. A higher completeness ensures that calibration is not dominated by a small local region, but rather benefits



(a) Mechanical LiDAR (b) Solid-state LiDAR (c) MEMS LiDAR (d) KITTI LiDAR

Fig. 8. Qualitative calibration results on outdoor scenes using four different LiDAR sensor types: (a) mechanical spinning LiDAR, (b) solid-state LiDAR, (c) MEMS LiDAR, and (d) Velodyne HDL-64E from KITTI. Despite variations in field of view, scanning patterns, and point density, calibrated point clouds align well with images, showing robustness of our method across different LiDAR types.

from globally balanced constraints, which improves both stability and accuracy.

Together, these two auxiliary metrics provide a more fine-grained perspective on how multi-view construction, geometric densification, and dual-channel matching contribute to the final calibration accuracy.

2) Ablation Configurations and Performance Comparison: We design six experimental configurations to separately evaluate the effects of multi-view construction, geometric densification, and dual-channel fusion on calibration accuracy and correspondence quality. The detailed results are summarized in Table IV. For consistency, the retention rate is computed with a distance threshold of 0.3 pixels, and completeness is evaluated using $K = 18$ virtual viewpoints.

3) Analysis and Discussion: As summarized in Table IV, and further illustrated by Fig. 9 and Fig. 10, the three proposed modules—multi-view construction, geometric densification, and dual-channel fusion—each contribute significantly to calibration robustness and accuracy.

Multi-view construction: As shown in Fig. 10, completeness steadily increases with the number of virtual viewpoints K, and begins to saturate around $K = 10$. Compared with the single-view case, introducing multiple viewpoints significantly improves the spatial balance of correspondences across the image, preventing calibration from being biased toward local regions. Table IV further confirms that multi-view projection yields lower rotation and translation errors.

Dual-channel fusion: As reported in Table IV, the dual-channel configuration (LIP+DEP) achieves a completeness of 81.64%, outperforming both single-channel variants (LIP: 68.16%, DEP: 75.49%). This confirms the complementary nature of intensity and depth, and shows that their fusion substantially increases the spatial coverage and robustness of cross-modal matching.

Geometric densification: Fig. 9 highlights the effect of densification on correspondence retention. For sparse point clouds, the retention rate remains near zero at small thresholds and increases only gradually with relaxed constraints. In contrast, densified point clouds exhibit an explosive growth in retention rate even at small thresholds, quickly achieving far higher match counts than sparse data.

Overall performance: In summary, multi-view construction improves coverage, densification enhances geometric constraints, and dual-channel fusion increases correspondence robustness. The increase in runtime mainly stems from the point cloud densification step, which introduces additional processing overhead, while the contributions of multi-view and dual-channel modules to runtime are relatively minor. Combined, these modules yield more accurate and stable calibration results.

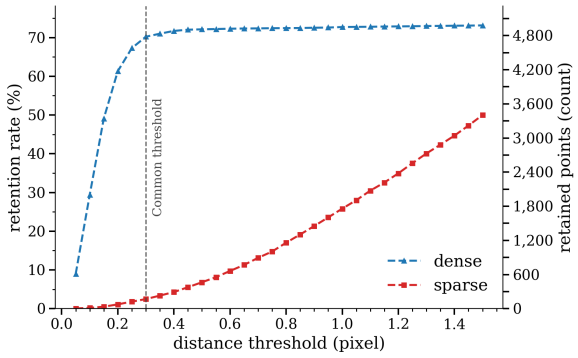


Fig. 9. Retention Rate vs. inlier distance threshold.

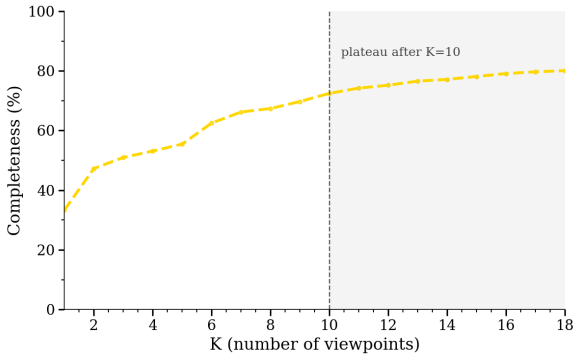


Fig. 10. Completeness vs. number of views K.

V. CONCLUSIONS

In this work, we presented MIND-Calib, a targetless LiDAR–camera calibration method that leverages virtual multi-view projection, dual-channel matching, and point cloud geometry densification. The approach requires only a single LiDAR–camera frame, achieves dense and consistent 2D–3D correspondences, and delivers accurate, robust performance across different LiDAR types and environments. Future directions include joint intrinsic–extrinsic optimization and embedded deployment.

TABLE IV

ABLATION STUDY ON DIFFERENT MODULES. EACH ABLATION IS EXECUTED WITH THE BEST CONFIGURATION IN OTHER PARAMETERS: THE VIEWPOINT ABLATION IS EXECUTED WITH DENSE AND DUAL-CHANNEL CONFIGURATION; THE CHANNEL ABLATION WITH MULTI-VIEW AND DENSE ; AND THE DENSIFICATION ABLATION WITH MULTI-VIEW AND DUAL-CHANNEL.

Module	Configuration	Errors		Retention Rate↑	Completeness↑	Runtime (s)↓
		e_R (deg)↓	e_t (cm)↓			
Viewpoint	Single-view	0.630	8.2	–	38.0	15.54
	Multi-view	**0.184**	**1.53**	–	**81.64**	48.83
Channel	LIP	0.680	6.6	–	68.16	29.13
	DEP	0.581	7.1	–	75.49	27.65
	LIP + DEP	**0.184**	**1.53**	–	**81.64**	48.83
Densification	Sparse	0.590	6.71	5.38	–	20.66
	Dense	**0.184**	**1.53**	**71.7**	–	48.83

ACKNOWLEDGMENT

We thank Prof. Guillermo Gallego (Technical University of Berlin) for his constructive comments that improved the clarity and presentation of this work. ChatGPT was used to provide suggestions for code optimization and to support diagram visualization design.

REFERENCES

[1] Y. Li, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, "ezfusion: A close look at the integration of lidar, millimeter-wave radar, and camera for accurate 3d object detection and tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 182–11 189, 2022.

[2] Z. Yu, W. Wan, M. Ren, X. Zheng, and Z. Fang, "Sparsefusion3d: Sparse sensor fusion for 3d object detection by radar and camera in environmental perception," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1524–1536, 2024.

[3] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 720–736.

[4] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for 3d detection and map segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 8721–8731.

[5] R. Ishikawa, S. Zhou, Y. Sato, T. Oishi, and K. Ikeuchi, "Lidar-camera calibration using intensity variance cost," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 10 688–10 694.

[6] Y. Zhu, C. Li, and Y. Zhang, "Online camera-lidar calibration with sensor semantic information," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4970–4976.

[7] X. Lv, B. Wang, Z. Dou, D. Ye, and S. Wang, "Lccnet: Lidar and camera self-calibration using cost volume network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 2894–2901.

[8] J. Zhu, J. Xue, and P. Zhang, "Calibdepth: Unifying depth map representation for iterative lidar-camera online calibration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 726–733.

[9] K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "General, single-shot, target-less, and automatic lidar-camera extrinsic calibration toolbox," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 301–11 307.

[10] Z. Huang, Y. Zhang, Q. Chen, and R. Fan, "Online, target-free lidar-camera extrinsic calibration via cross-modal mask matching," *IEEE Transactions on Intelligent Vehicles*, 2024.

[11] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight sam for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.

[12] J. Ren, X. Jiang, Z. Li, D. Liang, X. Zhou, and X. Bai, "Minima: Modality invariant image matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[13] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "RoMa: Robust Dense Feature Matching," *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[14] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the cpu," in *2018 15th Conference on Computer and Robot Vision (CRV)*, 2018, pp. 16–22.

[15] D. Teutscher, P. Mangat, and O. Wasenmüller, "Pdc: Piecewise depth completion utilizing superpixels," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2752–2758.

[16] Z. Pusztai and L. Hajder, "Accurate calibration of lidar-camera systems using ordinary boxes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[17] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, 2004, pp. 2301–2306 vol.3.

[18] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic targetless extrinsic calibration of a 3d lidar and camera by maximizing mutual information," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, p. 2053–2059.

[19] Z. Taylor and J. Nieto, "Automatic calibration of lidar and camera images using normalized mutual information," 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:22220583

[20] Z. Luo, G. Yan, X. Cai, and B. Shi, "Zero-training lidar-camera extrinsic calibration method using segment anything model," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 472–14 478.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3992–4003.

[22] P. Moghadam, M. Bosse, and R. Zlot, "Line-based extrinsic calibration of range and image sensors," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 3685–3691.

[23] C. Yuan, X. Liu, X. Hong, and F. Zhang, "Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7517–7524, 2021.

[24] D. Cattaneo and A. Valada, "Cmrnext: Camera to lidar matching in the wild for localization and extrinsic calibration," *IEEE Transactions on Robotics*, vol. 41, pp. 1995–2013, 2025.

[25] K. Petek, N. Vödisch, J. Meyer, D. Cattaneo, A. Valada, and W. Burgard, "Automatic target-less camera-lidar calibration from motion and deep point correspondences," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9978–9985, 2024.

[26] J. Tang, F.-P. Tian, B. An, J. Li, and P. Tan, "Bilateral propagation network for depth completion," *CVPR*, 2024.

[27] Y. Zhao, L. Bai, Z. Zhang, and X. Huang, "A surface geometry model for lidar depth completion," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4457–4464, 2021.

[28] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[29] A. Tsaregorodtsev, J. Muller, J. Strohbeck, M. Herrmann, M. Buchholz, and V. Belagiannis, "Extrinsic camera calibration with semantic segmentation," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 3781–3787.