This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3456299

Online, Target-Free LiDAR-Camera Extrinsic Calibration via Cross-Modal Mask Matching

Zhiwei Huang, *Graduate Student Member, IEEE*, Yikang Zhang, *Graduate Student Member, IEEE*, Qijun Chen, *Senior Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

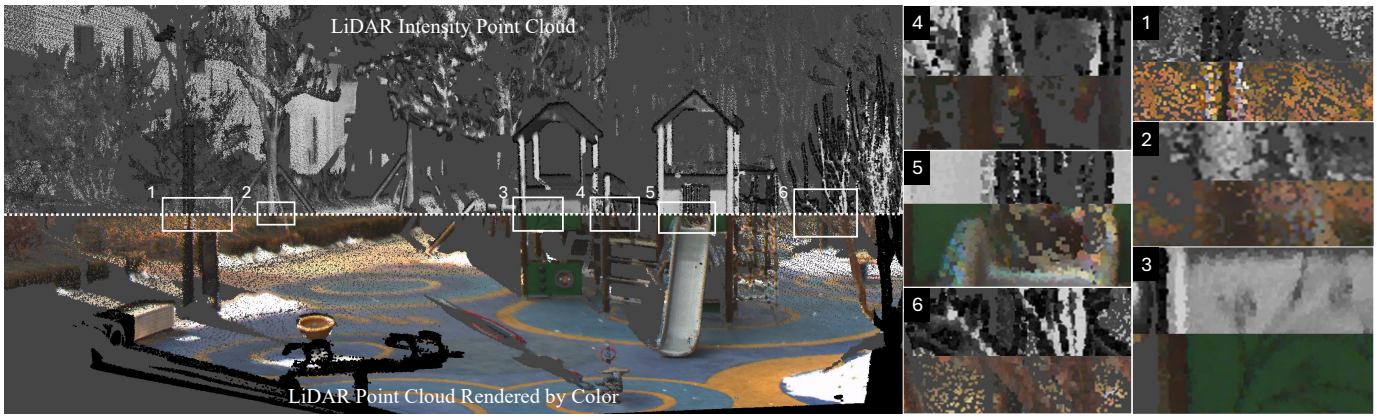


Fig. 1. Visualization of the experimental results achieved using our proposed online, target-free LCEC algorithm.

***Abstract*—LiDAR-camera extrinsic calibration (LCEC) is crucial for data fusion in intelligent vehicles. Offline, target-based approaches have long been the preferred choice in this field. 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. In contrast, online, target-free approaches provide greater adaptability yet typically lack robustness, primarily due to the challenges in cross-modal feature matching. Therefore, in this article, we unleash the full potential of large vision models (LVMs), which are emerging as a significant trend in the fields of computer vision and robotics, especially for embodied artificial intelligence, to achieve robust and accurate online, target-free LCEC across a variety of challenging scenarios. Our main contributions are threefold: we introduce a novel framework known as MIAS-LCEC, provide an open-source versatile calibration toolbox with an interactive visualization interface, and publish three real-world datasets captured from various indoor and outdoor environments. The cornerstone of our framework and toolbox is the cross-modal mask matching (C3M) algorithm, developed based on a state-of-the-art (SoTA) LVM and capable of generating sufficient and reliable matches. Extensive experiments conducted on these real-world datasets demonstrate the robustness of our approach and its superior performance compared to SoTA methods, particularly for the solid-state LiDARs with super-wide fields of view. Our toolbox and datasets are publicly available at https://mias.group/MIAS-LCEC.**

***Index Terms*—LiDAR-camera extrinsic calibration, intelligent vehicle, large vision model, embodied artificial intelligence.**

This research was supported by the National Science and Technology Major Project under Grant 2020AAA0108101, the National Natural Science Foundation of China under Grants 62473288 and 62233013, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program. (*Corresponding author: Rui Fan*)

Zhiwei Huang, Yikang Zhang, Qijun Chen, and Rui Fan are with the Department of Control Science & Engineering, the College of Electronics & Information Engineering, Shanghai Institute of Intelligent Science and Technology, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mails: {zhiweihuang, yikangzhang, qjchen}@tongji.edu.cn, rui.fan@ieee.org).

I. Introduction

A. Background

LIDARS provide accurate spatial geometric information, while cameras capture rich textural details [1]–[4]. Fusing data from both sensors enables intelligent vehicles to achieve more comprehensive 3D environmental perception [5]–[12]. LiDAR-camera extrinsic calibration (LCEC) forms the foundation for this data fusion process [13]–[16], as illustrated in Fig. 1. It basically estimates an extrinsic matrix ${}^{C}_{L}\boldsymbol{T}$, defined as follows [17]:

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

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. In this article, the symbols in the superscript and subscript denote the source and target sensors, respectively. When the camera intrinsic matrix $\boldsymbol{K}$ is known, a 3D LiDAR

point $\boldsymbol{p}^L = (x^L; y^L; z^L)$ can be projected onto a 2D image pixel $\boldsymbol{p} = (u; v)$ using the following expression:

$$\tilde{\boldsymbol{p}} = \frac{\boldsymbol{K}(\prescript{C}{L}{\boldsymbol{R}}\boldsymbol{p}^L + \prescript{C}{L}{\boldsymbol{t}})}{(\prescript{C}{L}{\boldsymbol{R}}\boldsymbol{p}^L + \prescript{C}{L}{\boldsymbol{t}})^\top \mathbf{1}_z}, \tag{2}$$

where $\tilde{\boldsymbol{p}}$ represents the homogeneous coordinates of $\boldsymbol{p}$ and $\mathbf{1}_z = (0; 0; 1)$. While extensive research on offline, target-based LCEC has yielded numerous effective and robust algorithms over decades, online, target-free methods, especially for solid-state LiDARs, remain less explored [18]. Consequently, this study aims to bridge this gap by leveraging state-of-the-art (SoTA) large vision models (LVMs).

B. Existing Challenges and Motivation

Existing online, target-free LCEC approaches first extract distinctive features, *e.g.*, line/edge features [18]–[23], point features [24], [25], or semantic features [26]–[30], from RGB images and LiDAR point clouds (spatial coordinates along with reflection intensities). These features are subsequently matched to produce cross-modal correspondences.

While line/edge feature-based LCEC approaches are highly efficient, their effectiveness is often limited by the requirement for sufficient and properly distributed edge features, confining their applicability to particular environments [18], [23]. First, edge detection algorithms typically identify straight lines, resulting in inadequate edge representations for objects with spherical or cylindrical geometries [20], [21]. Moreover, in scenarios where edges are predominantly aligned in one direction, the constraints may not be sufficient to uniquely determine the extrinsic parameters [19], [22]. Such scenarios can lead to solvers converging on local optima. Additionally, an uneven distribution of edges in an image can result in weak constraints that are susceptible to being influenced by measurement noise [21].

On the other hand, point feature-based LCEC approaches require distinctive 2D image pixels and 3D LiDAR points, characterized by significant changes in intensity or depth across all dimensions. However, this dependency may lead to a scarcity of viable correspondences, especially in low-texture environments [25]. Additionally, these methods often overlook the alignment of the field of view (FoV) between LiDAR and camera, resulting in excess of irrelevant points within the LiDAR point clouds, adversely impacting the LCEC process's stability [24].

Recent advances in deep learning techniques have spurred extensive exploration of semantic feature-based LCEC approaches. Although these approaches have shown compelling performance in specific scenarios, such as parking lots, they predominantly rely on curated, pre-defined objects, *e.g.*, vehicles [28], lanes [27], poles [26], and stop signs [30]. However, challenges such as domain shift (caused by differences in appearance, lighting conditions, or object distributions) and annotation inconsistency (different datasets often have diverse annotations) often impair the ability of these algorithms to generalize effectively across new, unseen scenarios.

Since 2023, LVMs have rapidly emerged as a focal point in the fields of computer vision and robotics. Models like Segmentation Anything (SAM) [31] and DINOv2 [32] by Meta AI, have attracted significant attention and interest due to their exceptional generalizability across new, complex, and challenging scenarios [33]. Therefore, this study aims to leverage SoTA LVMs to extract more informative features and develop a more robust cross-modal feature matching strategy, thereby improving the overall performance of online, target-free LCEC.

C. Novel Contributions

Therefore, in this article, we move one step forward in the field of online, target-free LCEC by unleashing the potential of MobileSAM [34], a SoTA LVM for image segmentation. First, we develop an online, target-free LCEC approach, referred to as MIAS-LCEC, which employs a novel coarse-to-fine strategy to accurately estimate LiDAR-camera extrinsic parameters. To minimize the modality discrepancy, we formulate the 3D-2D feature matching as a 2D-2D feature matching problem by introducing a virtual camera (whose pose is iteratively updated) to project the given LiDAR point cloud, thereby generating a LiDAR intensity projection (LIP) image, which appears as if it were taken from the perspective of the actual camera. This addresses the oversight of FoV alignment in the prior study [24] and helps achieve more effective and robust 3D-2D correspondence matching. Subsequently, both the LIP and RGB images undergo segmentation using MobileSAM. These segmentation results are then processed using a novel cross-modal mask matching (C3M) algorithm, capable of generating sparse yet reliable matches, which are propagated to target masks for dense matching. Finally, the obtained correspondences serve as inputs for a Perspective-n-Point (PnP) solver to derive the extrinsic matrix. Additionally, we launch a powerful toolbox with an interactive visualization interface. This toolbox also incorporates the manual calibration functionality, thereby further improving its utility. We collect three real-world datasets (from a variety of indoor and outdoor environments under various scenarios as well as different weather and illumination conditions) using a CMOS camera and different types of solid-state LiDARs to comprehensively evaluate the performance of LCEC algorithms. Through extensive experiments conducted on these datasets, our proposed MIAS-LCEC demonstrates superior robustness and accuracy compared to other SoTA online, target-free approaches. Moreover, it achieves a similar performance to that of an offline, target-based algorithm. Our toolbox and datasets are publicly available at https://mias.group/MIAS-LCEC.

In a nutshell, our main contributions are as follows:

- MIAS-LCEC, an online, target-free LCEC approach, which employs a novel coarse-to-fine strategy to accurately estimate LiDAR-camera extrinsic parameters by unleashing the potential of MobileSAM, a SoTA LVM for image segmentation.
- C3M, a novel and robust cross-modal feature matching algorithm, capable of generating dense and reliable correspondences.
- A versatile LCEC toolbox with an interactive visualization interface and capable of conducting online, target-free calibration and manual calibration.

- Three real-world datasets (containing dense 4D LiDAR point clouds and RGB images captured from a variety of indoor and outdoor environments), created to comprehensively evaluate the performance of LCEC algorithms.

D. Article Structure

The remainder of this article is structured as follows: Sect. II reviews SoTA approaches in LCEC. Sect. III introduces MIAS-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

Existing LCEC approaches are primarily categorized as either target-based or target-free based on whether the algorithm requires pre-defined features from both RGB images and LiDAR point clouds. The following two subsections discuss these two types of algorithms in detail.

A. Target-Based Approaches

The SoTA target-based LCEC approaches [35]–[40] are typically offline, relying on customized calibration targets (typically checkerboards). These targets enable the automatic detection of correspondences, as they provide distinct, recognizable features that can be easily identified in both LiDAR and camera data. For example, the study [35] performs extrinsic calibration by extracting corner points of a printed checkerboard from LiDAR point clouds and RGB images. It then optimizes the calibration result by formulating a RANSAC-based PnP problem, which minimizes the Euclidean distances between the corresponding corners. In the recent study [39], both intrinsic and extrinsic parameters are accurately estimated using a specially designed calibration target, which incorporates a checkerboard pattern and four specifically placed holes. While these methods achieve high calibration accuracy, their reliance on customized targets and the need for additional setup render it impractical for scenarios where robots operate in dynamically changing environments. Consequently, we introduce a fully target-free approach that provides greater applicability and flexibility, eliminating the need for specialized calibration targets. This approach enables robots to rapidly obtain high-precision extrinsic parameters anytime and anywhere.

B. Target-Free Approaches

To improve the environmental adaptability of LCEC, previous studies [41]–[43] have shifted from relying on specific targets to extracting informative visual features directly from the environment. In early attempts, researchers manually identified cross-modal correspondences and conducted LCEC using the PnP pose estimation algorithm [41], [42]. Nevertheless, this manual LCEC process is tedious and prone to errors introduced by the manually selected correspondences [18].

Afterwards, traditional line/edge feature-based automatic LCEC approaches emerged. In studies such as [19], [22], LiDAR point intensities are first projected into the camera perspective, thereby generating a virtual image, namely an LIP image. Edges are then extracted from both the LIP and RGB images. By matching these cross-modal edges, the relative pose between the two sensors can be determined. Similarly, research by [44], [45] optimizes extrinsic calibration by maximizing the mutual information (MI) between LIP and RGB images. To address occlusion issues in [44], [45], the study of HKU-Mars [21] employs a voxelization method to detect and extract 3D lines from the point cloud, achieving high accuracy in scenarios with rich 3D line features. While effective in specific scenarios with abundant features, these traditional methods heavily rely on well-distributed line/edge features, which can compromise calibration robustness. Moreover, the use of low-level image processing algorithms, such as Gaussian blur and the Canny operator, can introduce errors in edge detection, potentially fragmenting global lines and thus reducing overall calibration accuracy.

Advances in deep learning techniques have driven significant exploration into enhancing traditional line/edge feature-based algorithms. In [18], edge detection with Transformer (EDTER) [46], a deep neural network, is utilized to improve the accuracy of 2D edge detection in RGB images. Additionally, a supervoxel-based 3D line detection method was designed to detect global line features from 3D point clouds, improving the line-based method proposed in [21]. Despite its impressive performance, this approach still heavily relies on specific scenarios with abundant properly-distributed lines, ultimately limiting its applicability.

To overcome this limitation, several end-to-end deep learning-based algorithms [47]–[49] have been developed. RegNet [47], the first convolutional neural network, developed specifically for extrinsic parameter estimation, is trained by minimizing a loss function representing the distance between predicted and ground-truth parameters. However, RegNet requires retraining when sensor intrinsic parameters change. CalibNet [48] improves RegNet by maximizing geometric and photometric consistency between point clouds and images, thereby regressing extrinsic parameters implicitly. Another end-to-end network, LCCNet [49], introduces a cross-attention module to measure the similarity between point clouds and images. While these methods have demonstrated effectiveness on large-scale datasets like KITTI [50], which primarily focuses on urban driving scenarios, their performance has not been extensively validated on other types of real-world datasets. Furthermore, their dependence on pre-defined sensor configurations (both LiDAR and camera) poses implementation challenges.

Inspired by end-to-end keypoint detection and matching neural networks, a recent study [24] introduced Direct Visual LiDAR Calibration (DVL), a novel point-based method that utilizes SuperGlue [51] to establish direct 3D-2D correspondences between LiDAR and camera data. Additionally, this study refines the estimated extrinsic matrix through direct LiDAR-camera registration by minimizing the normalized information distance, a mutual information-based cross-modal distance measurement loss. However, an oversight in aligning the FoV between the LiDAR and camera leads to the presence

of numerous redundant points of interest within the LiDAR point clouds, adversely affecting the overall stability of the LCEC process. Therefore, in this article, we improve the LiDAR point intensity projection strategy to generate LIP images that are more similar to the RGB images.

To further improve the robustness of cross-modal feature matching for LCEC, several studies [13], [26]–[30] have resorted to semantic segmentation techniques. For instance, in [28], parking vehicles are first detected and then used to register point clouds with images. The study [13] maximizes the overlapping area of vehicles as represented in both point clouds and images. Similarly, [27] accomplishes LiDAR and camera registration by aligning road lanes and poles, while [30] employs stop signs as calibration primitives and refines results over time using a Kalman filter. Moreover, [29] investigates the consistency of segmented edges between point clouds and images, and [26] introduces an automatic registration method based on pole matching. However, as discussed earlier, domain shift and annotation inconsistency issues often hinder these algorithms from effectively generalizing to unseen, new scenarios. In this work, we take a pioneering step by leveraging SoTA LVMs to extract more informative features from both LIP and RGB images. Furthermore, we develop a more robust cross-modal feature matching strategy, which makes the LCEC process fully target-free, overcoming the previous reliance on specific semantic targets. By integrating these advancements, we aim to enhance the accuracy and robustness of LCEC algorithms, enabling them to perform effectively in diverse and challenging real-world environments.

III. Methodology

A. Algorithm Overview

Our proposed online, target-free LCEC algorithm MIAS-LCEC, as depicted in Fig. 2, employs a novel coarse-to-fine pipeline. A virtual camera projects LiDAR point intensities into the camera perspective. Both the resulting LIP image and the RGB image are processed using MobileSAM, a SoTA LVM for image segmentation. Sufficient and reliable correspondences identified by our C3M strategy are then used as inputs for a PnP solver to estimate the extrinsic matrix ${}^{C}_{L}\boldsymbol{T}$.

Previous studies [24], [52] typically set up the virtual camera with a relative transformation ${}^{V}_{L}\boldsymbol{T}$ from LiDAR as follows:

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

thereby generating an LIP image $\boldsymbol{I}^{L} \in \mathbb{R}^{H \times W \times 1}$ to formulate the LCEC problem as a 2D feature matching problem, where H and W denote its height and width, respectively. Considering the image distortion introduced by different perspective views, (3) constrains the sensor setup to a captious relative transformation. Therefore, we propose a novel strategy to iteratively refine the virtual camera pose until the LIP image resembles one taken from the actual camera's perspective view. This iterative process can be expressed as follows:

$$ \lim_{k \to +\infty} {}^{C}_{V}\boldsymbol{T}_{k} = \lim_{k \to +\infty} ({}^{V}_{L}\boldsymbol{T}_{k})^{-1} {}^{C}_{L}\boldsymbol{T} \approx \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0}^{\top} & 1 \end{pmatrix}, \tag{4} $$

where the subscript k denotes the k-th iteration, ${}^{V}_{L}\boldsymbol{T}_{k}$ represents the transformation from LiDAR to the virtual camera, and $\boldsymbol{I}$ denotes the identity matrix.

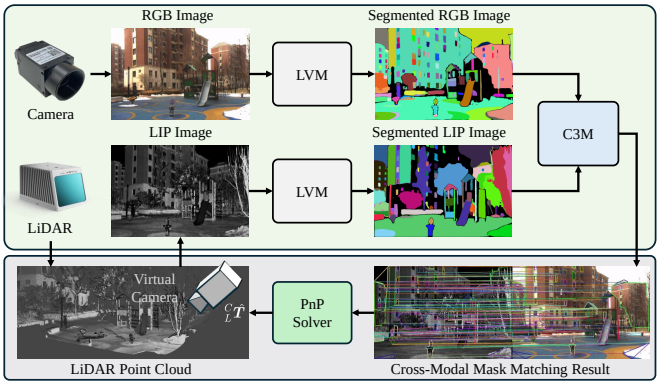


Fig. 2. The pipeline of our proposed online, target-free LCEC algorithm.

Using the LIP image captured in each iteration and our proposed C3M in Sect. III-B, we can generate two sets $\mathcal{P}_k = \{\boldsymbol{p}_{k,1}, \ldots, \boldsymbol{p}_{k,N_k}\}$ and $\mathcal{P}^{L}_{k} = \{\boldsymbol{p}^{L}_{k,1}, \ldots, \boldsymbol{p}^{L}_{k,N_k}\}$, which store 2D pixels in the RGB image captured by camera and the corresponding 3D LiDAR points, respectively. The extrinsic matrix ${}^{C}_{L}\hat{\boldsymbol{T}}_{k}$ can then be effectively computed by minimizing the mean reprojection error as follows:

$$ {}^{C}_{L}\hat{\boldsymbol{T}}_{k} = \underset{{}^{C}_{L}\boldsymbol{T}_{k,i}}{\arg\min} \underbrace{\frac{1}{N_k} \sum_{n=1}^{N_k} \left\| \boldsymbol{K}({}^{C}_{L}\boldsymbol{R}_{k,i}\boldsymbol{p}^{L}_{k,n} + {}^{C}_{L}\boldsymbol{t}_{k,i}) - \boldsymbol{p}_{k,n} \right\|_2}_{\epsilon_k}, \tag{5} $$

where ${}^{C}_{L}\boldsymbol{T}_{k,i} = \begin{pmatrix} {}^{C}_{L}\boldsymbol{R}_{k,i} & {}^{C}_{L}\boldsymbol{t}_{k,i} \\ \boldsymbol{0}^{\top} & 1 \end{pmatrix} \in SE(3)$ denotes the i-th PnP solution obtained using a selected subset of correspondences from $\mathcal{P}_k$ and $\mathcal{P}^{L}_{k}$, and ϵ_k represents the mean reprojection error with respect to ${}^{C}_{L}\boldsymbol{T}_{k,i}$.

Our MIAS-LCEC algorithm updates ${}^{V}_{L}\boldsymbol{T}_{k+1}$ with ${}^{C}_{L}\hat{\boldsymbol{T}}_{k}$. According to (4), ${}^{V}_{L}\boldsymbol{T}_{k+1} = {}^{C}_{L}\hat{\boldsymbol{T}}_{k} \approx {}^{C}_{L}\boldsymbol{T}$ as the iterative process converges, minimizing the calibration error to the greatest extent. In practical applications, to optimize the trade-off between accuracy and efficiency, we terminate the iterative process when $\epsilon_{k+1} > \epsilon_k$, and select ${}^{C}_{L}\hat{\boldsymbol{T}}_{k}$ from the k-th iteration as the final calibration result, namely ${}^{C}_{L}\boldsymbol{T}^{*}$.

B. Cross-Modal Mask Matching

In this article, we adopt a two-stage strategy to realize cross-modal mask matching, as detailed in Algorithm 1. Each stage consists of sequential coarse instance matching and fine-grained corner point matching. The first stage yields reliable yet sparse matches, from which we derive the parameters for an affine transformation to update the masks within the LIP image. In the second stage, we achieve dense mask matching by propagating the obtained reliable reference matches to the

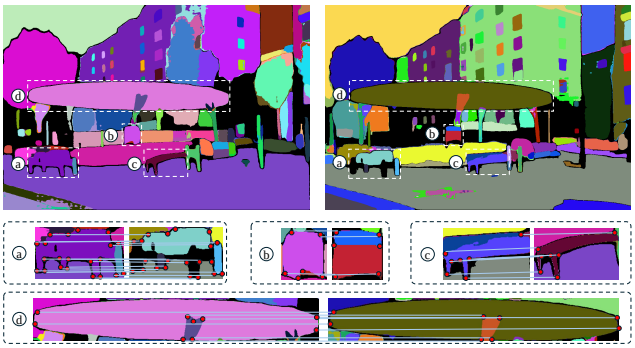


Fig. 3. An example of two-stage, coarse-to-fine cross-modal mask matching result: (a)-(d) illustrate four examples of instance matching and corner point matching results. Potential errors produced by the LVM are greatly minimized through our strict match selection criterion.

target masks. These dense matches are finally utilized as inputs for the PnP solver to obtain ${}^{C}_{L}\boldsymbol{T}$.

As shown in Fig. 3, our developed C3M strategy significantly minimizes potential matching errors produced by the LVM, primarily due to the strict match selection criterion. The corner points along the contours of masks detected within the LIP image and the RGB image using Mobile-SAM are represented by two sets: $\mathcal{C}^V = \{\boldsymbol{c}_1^V, \ldots, \boldsymbol{c}_m^V\}$ and $\mathcal{C}^C = \{\boldsymbol{c}_1^C, \ldots, \boldsymbol{c}_m^C\}$, respectively. An instance (bounding box), utilized to precisely fit around each mask, is centrally positioned at $\boldsymbol{o}^{V,C}$ and has a dimension of $h^{V,C} \times w^{V,C}$ pixels. To determine optimum instance matches, we construct a cost matrix $\boldsymbol{M}^I$, where the element at $\boldsymbol{x} = (i; j)$, namely:

$$\boldsymbol{M}^I(\boldsymbol{x}) = \frac{1}{4}\left(\frac{|w^C - w^V|}{w^C + w^V} + \frac{|h^C - h^V|}{h^C + h^V} + 2\left(1 - \exp\left(-\frac{\left\|\hat{\boldsymbol{o}}^V - \boldsymbol{o}^C\right\|_2}{h^C + h^V + w^C + w^V}\right)\right)\right) \in [0, 1], \tag{6}$$

denotes the matching cost between the i-th instance from the LIP image and the j-th instance from the RGB image. $\hat{\boldsymbol{o}}^V$ is initially set as $\boldsymbol{o}^V$ during the sparse matching phase and subsequently updated using the above-mentioned affine transformation prior to the dense matching phase, as illustrated in Fig. 4, so as to minimize the discrepancies arising from the differing perspectives between LiDAR and camera. A strict criterion is applied to achieve sparse yet reliable matching. Matches with the lowest costs in both horizontal and vertical directions are determined as the optimum coarse instance matches.

Subsequently, we determine corner point correspondences within the matched instances. Similarly, a cost matrix $\boldsymbol{M}^C$ is constructed, where the element at $\boldsymbol{y} = (r; s)$, namely:

$$\boldsymbol{M}^C(\boldsymbol{y}) = \frac{\|(\hat{\boldsymbol{c}}^V - \hat{\boldsymbol{o}}^V) - (\boldsymbol{c}^C - \boldsymbol{o}^C)\|_2}{\|(\hat{\boldsymbol{c}}^V - \hat{\boldsymbol{o}}^V)\|_2 + \|(\boldsymbol{c}^C - \boldsymbol{o}^C)\|_2} \in [0, 1], \tag{7}$$

denotes the matching cost between the r-th corner point of a mask in the LIP image and the s-th corner point of a mask in the RGB image. $\hat{\boldsymbol{c}}^V$ is initialized as $\boldsymbol{c}^V$ during the sparse matching phase and updated using the same affine transformation prior to the dense matching phase. Correspondences with the lowest costs both horizontally and vertically are also determined to be optimum corner point matching results. Nevertheless, the first stage is considerably critical and cannot often provide the PnP solver with sufficient inputs.

Therefore, we apply an affine transformation to the masks within the LIP image to adjust $\boldsymbol{o}^V$ and $\boldsymbol{c}^V$, as follows:

$$\begin{cases} \hat{\boldsymbol{c}}^V = s\boldsymbol{R}^A\boldsymbol{c}^V + \boldsymbol{t}^A \\ \hat{\boldsymbol{o}}^V = s\boldsymbol{R}^A\boldsymbol{o}^V + \boldsymbol{t}^A \end{cases}, \tag{8}$$

where $\boldsymbol{R}^A \in SO(2)$ represents the rotation matrix, $\boldsymbol{t}^A$ denotes the translation vector, and s represents the scaling factor. Given

Algorithm 1 Cross-Modal Mask Matching

Require:
Cross-modal masks, obtained from the LIP and RGB images.
Stage 1 (Reliable sparse matching):
(1) Construct the instance matching cost matrix $\boldsymbol{M}^I$ using (6).
(2) Select matched instances with low costs from $\boldsymbol{M}^I$ as reliable matches.
(3) Construct $\boldsymbol{M}^C$ using (7) and match corner points.
(4) Estimate $s\boldsymbol{R}^A$ and $\boldsymbol{t}^A$ using (9)-(12).
Stage 2 (Dense mask matching):
(1) Update all masks in the LIP image using $s\boldsymbol{R}^A$ and $\boldsymbol{t}^A$.
(2) Update $\boldsymbol{M}^I$ with the updated masks to obtain dense instance matching results.
(3) For each pair of matched instances, update their $\boldsymbol{M}^C$ using (7) to determine corner point correspondences.
(4) Aggregate all corner point correspondences to form the sets $\mathcal{P}$ and $\mathcal{P}^L$.

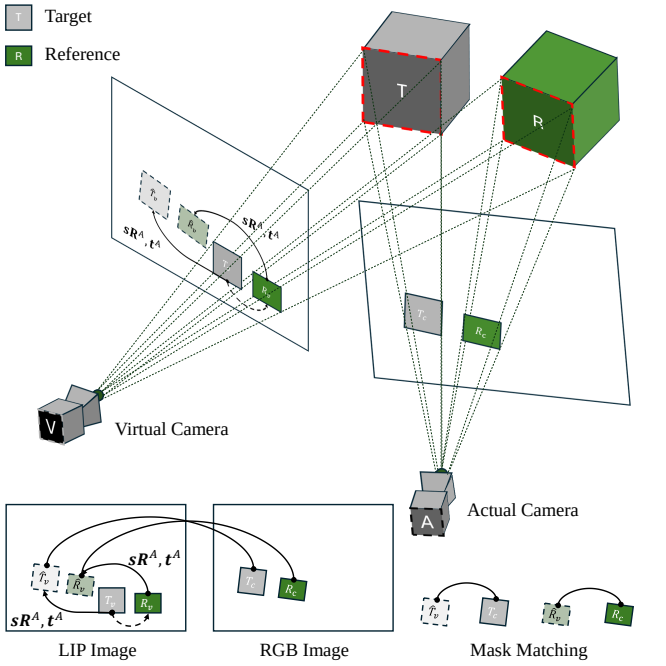


Fig. 4. Upon matching the target masks T_v and T_c, an affine transformation estimated from a pair of reference masks R_v and R_c, is used to update the mask in the LIP image, so as to more accurately reflect the actual matching relationship.

the critical nature of our designed sparse matching strategy, we can assume that after the affine transformation, any points within a given mask in the LIP image perfectly align with the corresponding points from the RGB image, and thus, $\hat{\boldsymbol{c}}^V = \boldsymbol{c}^C$ and $\hat{\boldsymbol{o}}^V = \boldsymbol{o}^C$. In this case, $\boldsymbol{R}^A$ can be obtained using the following expression:

$$\boldsymbol{R}^A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{9}$$

where

$$\theta = \frac{1}{N}\sum_{i=1}^{N}\left(\arctan\frac{\mathbf{1}_y^\top(\boldsymbol{c}_i^V - \boldsymbol{o}^V)}{\mathbf{1}_x^\top(\boldsymbol{c}_i^V - \boldsymbol{o}^V)} - \arctan\frac{\mathbf{1}_y^\top(\boldsymbol{c}_i^C - \boldsymbol{o}^C)}{\mathbf{1}_x^\top(\boldsymbol{c}_i^C - \boldsymbol{o}^C)}\right) \tag{10}$$

is the angle between the vectors originating from the mask centers and pointing to their respective matched corner points. s can then be expressed as follows:

$$s = \frac{w^C h^C}{w^V h^V}, \tag{11}$$

which represents the ratio between the areas of the bounding boxes associated with the RGB image and the LIP image. Finally, according to (8), $\boldsymbol{t}^A$ can be obtained as follows:

$$\boldsymbol{t}^A = \boldsymbol{o}^C - s\boldsymbol{R}^A\boldsymbol{o}^V. \tag{12}$$

The remainder of this subsection delves into the relationship between a given pair of matched corner points within the reference and target masks, demonstrating the feasibility and reasonableness of propagating sparse, reliable mask matches.

Given two 3D LiDAR points $\boldsymbol{q}^V$ (reference) and $\boldsymbol{p}^V$ (target) in the virtual camera coordinate system, their correspondences, $\boldsymbol{q}^C$ and $\boldsymbol{p}^C$, in the actual camera coordinate system can be established through the following transformations:

$$\boldsymbol{q}^C = {}^C_V\boldsymbol{R}\boldsymbol{q}^V + {}^C_V\boldsymbol{t}, \tag{13}$$

$$\boldsymbol{p}^C = {}^C_V\boldsymbol{R}\boldsymbol{p}^V + {}^C_V\boldsymbol{t}. \tag{14}$$

$\tilde{\boldsymbol{q}}_{v,c}$ and $\tilde{\boldsymbol{p}}_{v,c}$, the homogeneous coordinates of the corresponding 2D pixels of $\boldsymbol{q}^{V,C}$ and $\boldsymbol{p}^{V,C}$ in LIP and RGB images, can be obtained as follows:

$$\begin{cases} \tilde{\boldsymbol{p}}_v = \dfrac{\boldsymbol{K}}{(\boldsymbol{p}^V)^\top\mathbf{1}_z}\boldsymbol{p}^V, & \tilde{\boldsymbol{p}}_c = \dfrac{\boldsymbol{K}}{(\boldsymbol{p}^C)^\top\mathbf{1}_z}\boldsymbol{p}^C, \\ \tilde{\boldsymbol{q}}_v = \dfrac{\boldsymbol{K}}{(\boldsymbol{q}^V)^\top\mathbf{1}_z}\boldsymbol{q}^V, & \tilde{\boldsymbol{q}}_c = \dfrac{\boldsymbol{K}}{(\boldsymbol{q}^C)^\top\mathbf{1}_z}\boldsymbol{q}^C. \end{cases} \tag{15}$$

Plugging (15) into (13) results in the affine transformation from $\tilde{\boldsymbol{q}}_v$ to $\tilde{\boldsymbol{q}}_c$ as follows:

$$\tilde{\boldsymbol{q}}_c = \underbrace{\frac{(\boldsymbol{q}^V)^\top\mathbf{1}_z}{(\boldsymbol{q}^C)^\top\mathbf{1}_z}\boldsymbol{K}({}^C_V\boldsymbol{R})\boldsymbol{K}^{-1}}_{\boldsymbol{A}}\tilde{\boldsymbol{q}}_v + \underbrace{\frac{\boldsymbol{K}}{(\boldsymbol{q}^C)^\top\mathbf{1}_z}{}^C_V\boldsymbol{t}}_{\boldsymbol{b}}. \tag{16}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ represent an affine transformation from $\boldsymbol{q}_v$ to $\boldsymbol{q}_c$. Combining (14) and (13) results in the following expression:

$$\boldsymbol{p}^C = \boldsymbol{q}^C + {}^C_V\boldsymbol{R}(\boldsymbol{p}^V - \boldsymbol{q}^V). \tag{17}$$

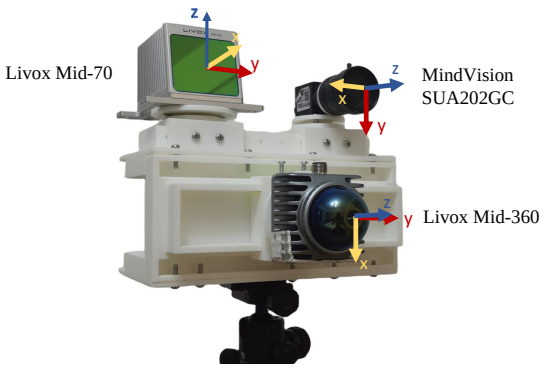


Fig. 5. Our experimental setup, where two solid-state Livox LiDARs and one MindVision camera are utilized for data acquisition.

Plugging (15) into (17) results in the following expression:

$$\begin{aligned} \tilde{\boldsymbol{p}}_c =& \frac{(\boldsymbol{q}^C)^\top\mathbf{1}_z}{(\boldsymbol{p}^C)^\top\mathbf{1}_z}\tilde{\boldsymbol{q}}_c \\ &+ \frac{(\boldsymbol{p}^V)^\top\mathbf{1}_z}{(\boldsymbol{p}^C)^\top\mathbf{1}_z}\boldsymbol{K}({}^C_V\boldsymbol{R})\boldsymbol{K}^{-1}(\tilde{\boldsymbol{p}}_v - \frac{(\boldsymbol{q}^V)^\top\mathbf{1}_z}{(\boldsymbol{p}^V)^\top\mathbf{1}_z}\tilde{\boldsymbol{q}}_v) \\ =& \frac{(\boldsymbol{p}^V)^\top\mathbf{1}_z}{(\boldsymbol{p}^C)^\top\mathbf{1}_z}\frac{(\boldsymbol{q}^C)^\top\mathbf{1}_z}{(\boldsymbol{q}^V)^\top\mathbf{1}_z}\underbrace{\frac{(\boldsymbol{q}^V)^\top\mathbf{1}_z}{(\boldsymbol{q}^C)^\top\mathbf{1}_z}\boldsymbol{K}({}^C_V\boldsymbol{R})\boldsymbol{K}^{-1}}_{\boldsymbol{A}}\tilde{\boldsymbol{p}}_v \\ &+ \frac{(\boldsymbol{q}^C)^\top\mathbf{1}_z}{(\boldsymbol{p}^C)^\top\mathbf{1}_z}\underbrace{\left(\tilde{\boldsymbol{q}}_c - \frac{(\boldsymbol{q}^V)^\top\mathbf{1}_z}{(\boldsymbol{q}^C)^\top\mathbf{1}_z}\boldsymbol{K}({}^C_V\boldsymbol{R})\boldsymbol{K}^{-1}\tilde{\boldsymbol{q}}_v\right)}_{\boldsymbol{b}}. \end{aligned} \tag{18}$$

When $\boldsymbol{p}^V$ and $\boldsymbol{q}^V$ are close in depth, namely $(\boldsymbol{p}^V)^\top\mathbf{1}_z \approx (\boldsymbol{p}^C)^\top\mathbf{1}_z \approx (\boldsymbol{q}^V)^\top\mathbf{1}_z \approx (\boldsymbol{q}^C)^\top\mathbf{1}_z$, (18) can be rewritten as follows:

$$\tilde{\boldsymbol{p}}_c \approx \boldsymbol{A}\tilde{\boldsymbol{p}}_v + \boldsymbol{b}. \tag{19}$$

(16) and (19) indicate that $\tilde{\boldsymbol{q}}_{v,c}$ and $\tilde{\boldsymbol{p}}_{v,c}$ can share the same affine transformation when $\boldsymbol{q}^{V,C}$ and $\boldsymbol{p}^{V,C}$ are close in depth. In practice, we use the following affine transformation:

$$\begin{pmatrix} \tilde{\boldsymbol{p}}_c & \tilde{\boldsymbol{q}}_c \end{pmatrix} = \begin{pmatrix} s\boldsymbol{R}^A & \boldsymbol{t}^A \\ \mathbf{0}^\top & 1 \end{pmatrix}\begin{pmatrix} \tilde{\boldsymbol{p}}_v & \tilde{\boldsymbol{q}}_v \end{pmatrix}. \tag{20}$$

Therefore, the affine transformation of the target mask can be approximated by $s\boldsymbol{R}^A$ and $\boldsymbol{t}^A$, which are derived from the reference masks that are spatially close to the target mask.

In the first stage of the C3M process, reference masks are not yet identified. Therefore, $s\boldsymbol{R}^A$ and $\boldsymbol{t}^A$ are not considered when constructing $\boldsymbol{M}^I$ and $\boldsymbol{M}^C$, and are initialized as $s\boldsymbol{R}^A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\boldsymbol{t}^A = (0;0)$. In the second stage, they are determined using (9)-(12), based on reliable reference masks identified in the first stage. The entire C3M process is efficient because it primarily focuses on 2D affine transformation, rather than complex 3D point cloud registration.

IV. Experiment

A. Experimental Setup

Our experimental setup, as shown in Fig. 5, consisting of two solid-state Livox LiDARs (Mid-70 and Mid-360) from

DJI and an MV-SUA202GC global-shutter CMOS camera from MindVision, is used for cross-modal data collection. The Mid-70 and Mid-360 LiDARs both operate at a point rate of 200,000 points/s. However, the Mid-70 LiDAR captures dual returns, while the Mid-360 LiDAR captures only the first return. The RGB image resolution is 1,200×800 pixels. The camera's intrinsic parameters are determined through offline calibration prior to the experiments and are presumed to remain constant.

We compare our method with four SoTA target-free LCEC approaches: CRLF [27], UMich [45], HKU-Mars [21] and DVL [24], on two real-world datasets, MIAS-LCEC-TF70 and MIAS-LCEC-TF360. Additionally, to validate the effectiveness of our method in scenarios where targets are present, we also conduct comparisons with a classical offline target-based LCEC approach introduced in [35] on the MIAS-LCEC-CB70 dataset.

Our algorithm was implemented on an Intel i7-14700K CPU and an NVIDIA RTX4070Ti Super GPU. The entire process, including data preprocessing, C3M, and extrinsic parameters optimization, takes approximately 15 to 70 seconds.

B. Datasets

We have created the following three real-world datasets: **MIAS-LCEC-TF70** (target-free), **MIAS-LCEC-CB70** (target-based), and **MIAS-LCEC-TF360** (target-free), which are now publicly available for researchers to evaluate the performance of LCEC approaches:

- MIAS-LCEC-TF70 is a diverse and challenging dataset that contains 60 pairs of 4D point clouds (including spatial coordinates with intensity data) and RGB images, collected using a Livox Mid-70 LiDAR and a MindVision SUA202GC camera, from a variety of indoor and outdoor environments, under various scenarios as well as different weather and illumination conditions. We divide this dataset into six subsets: residential community, urban freeway, building, challenging weather, indoor, and challenging illumination, to comprehensively evaluate the algorithm performance.
- MIAS-LCEC-CB70 contains 15 pairs of 4D point clouds and RGB images, all collected in our laboratory using a Livox Mid-70 LiDAR and a MindVision SUA202GC camera, with yaw angles ranging from -30° to +30°, and the distances between the sensors and a calibration checkerboard pattern ranging from 3 m to 5 m. The checkerboard pattern comprises alternating white and black squares of equal size (8 cm × 8 cm).
- MIAS-LCEC-TF360 contains 12 pairs of 4D point clouds and RGB images, collected using a Livox Mid-360 LiDAR and a MindVision SUA202GC camera from both indoor and outdoor environments. Since the Livox Mid-360 LiDAR has a scanning range of 360°, it produces a sparser point cloud compared to that generated by the Livox Mid-70 LiDAR. Additionally, the significant difference in the FoV between this type of LiDAR and the camera results in only a small overlap in the collected data. Consequently, this dataset is particularly well-suited for evaluating the adaptability of algorithms to challenging scenarios characterized by sparse point clouds and limited data overlap.

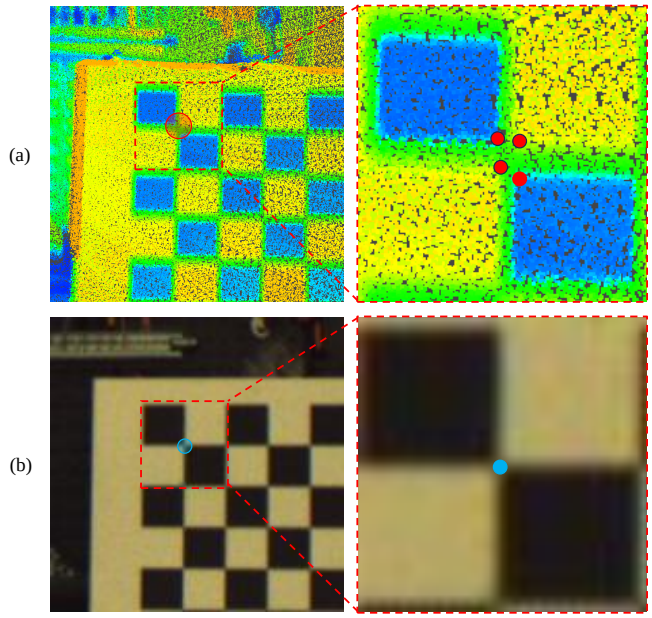


Fig. 6. Comparison of corner point detection between LiDAR point cloud and RGB image: (a) possible detection results in the LiDAR point cloud; (b) the detection result in the RGB image. The blue point in (b) has four possible matches in (a).

C. Evaluation Metrics

In our experiments, the Euler angle error, with the following expression:

$$e_r = \|\boldsymbol{r}^* - \boldsymbol{r}\|_2 , \tag{21}$$

where $\boldsymbol{r}^*$ and $\boldsymbol{r}$ represent the estimated and ground-truth Euler angle vectors, computed from the rotation matrices ${}^C_L\boldsymbol{R}^*$ and ${}^C_L\boldsymbol{R}$, respectively, and the translation error, with the following expression[1]:

$$e_t = \left\| -({}^C_L\boldsymbol{R}^*)^{-1}\boldsymbol{t}^* + {}^C_L\boldsymbol{R}^{-1}\boldsymbol{t} \right\|_2 , \tag{22}$$

where $\boldsymbol{t}^*$ and $\boldsymbol{t}$ denote the estimated and ground-truth translation vectors, respectively, are used to quantify the performance of target-free LCEC approaches.

Additionally, we use the following reprojection error

$$\epsilon = \frac{1}{M}\sum_{i=1}^{M} \left\| \boldsymbol{K}({}^C_L\boldsymbol{R}^*\boldsymbol{p}_i^L + {}^C_L\boldsymbol{t}^*) - \tilde{\boldsymbol{p}}_i \right\|_2 . \tag{23}$$

between 3D LiDAR points and their corresponding 2D image pixels to quantify the performance of LCEC algorithms when using targets.

As illustrated in Fig. 6, we observe that LiDAR scanning near textural and geometric discontinuities typically exhibits inherent errors, resulting in a reprojection error of around one pixel. Therefore, in our experiments, we consider results with a reprojection error of less than two pixels to be satisfactory.

[1]The translation from LiDAR pose to camera pose is $-{}^C_L\boldsymbol{R}^{-1}\boldsymbol{t}$ when (1) is used to depict the point translation.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3456299

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

Subsets	Approach	Rotation Error (°)						Translation Error (m)					
		Yaw	Pitch	Roll	e_r (°)			X	Y	Z	e_t (m)		
					Mean	Max	Min				Mean	Max	Min
Residential Community	CRLF [27]	0.177	1.581	0.876	1.594	1.754	1.537	0.162	0.429	0.060	0.464	3.636	0.136
	UMich [45]	0.690	1.502	4.066	4.829	21.706	0.214	0.306	0.151	0.058	0.387	2.394	0.072
	HKU-Mars [21]	1.704	1.128	1.113	2.695	7.036	0.749	0.829	0.458	0.397	1.208	4.405	0.319
	DVL [24]	0.107	**0.089**	**0.091**	0.193	**0.383**	0.042	0.034	0.043	**0.018**	0.063	**0.141**	0.018
	MIAS-LCEC (Ours)	**0.049**	0.138	0.094	**0.190**	0.711	**0.012**	**0.032**	**0.016**	0.026	**0.050**	0.176	**0.018**
Urban Freeway	CRLF [27]	0.434	1.584	0.857	1.582	1.585	1.581	0.043	0.130	0.030	0.140	0.140	0.140
	UMich [45]	0.471	1.331	1.464	2.267	6.531	0.506	**0.051**	0.128	0.063	0.166	0.241	0.103
	HKU-Mars [21]	2.518	1.610	0.955	2.399	3.744	0.337	1.482	1.125	0.284	1.956	9.453	0.148
	DVL [24]	**0.212**	0.140	**0.045**	0.298	**0.420**	0.090	0.103	0.056	0.017	0.124	0.268	0.064
	MIAS-LCEC (Ours)	0.216	**0.111**	0.083	**0.291**	0.514	**0.062**	0.070	**0.052**	**0.010**	**0.111**	**0.145**	**0.047**
Building	CRLF [27]	0.231	1.409	0.911	1.499	1.706	1.465	17.078	8.692	6.168	20.165	140.316	0.140
	UMich [45]	0.927	1.824	15.097	11.914	22.778	0.452	0.544	0.182	0.407	0.781	1.497	0.042
	HKU-Mars [21]	0.800	1.391	0.589	1.814	3.618	0.118	0.532	0.262	0.227	0.706	2.595	0.059
	DVL [24]	0.138	**0.103**	**0.049**	0.200	0.357	0.059	0.058	0.047	0.031	0.087	0.146	0.052
	MIAS-LCEC (Ours)	**0.080**	0.152	0.074	**0.198**	**0.390**	**0.051**	**0.049**	**0.036**	**0.020**	**0.072**	**0.137**	**0.024**
Challenging Weather	CRLF [27]	0.183	1.647	0.917	1.646	1.803	1.552	1.614	0.867	0.617	2.055	20.307	0.137
	UMich [45]	0.383	1.635	3.720	1.851	5.335	0.294	0.212	0.109	0.103	0.310	2.134	0.043
	HKU-Mars [21]	1.110	1.128	1.840	2.578	11.647	0.371	0.772	0.456	0.469	1.086	4.934	0.302
	DVL [24]	**0.095**	**0.100**	0.069	0.181	**0.311**	0.037	0.029	0.036	**0.015**	0.052	**0.113**	**0.017**
	MIAS-LCEC (Ours)	0.100	0.112	**0.059**	**0.177**	0.412	**0.030**	**0.027**	**0.027**	0.018	**0.046**	0.127	0.021
Indoor	CRLF [27]	0.164	1.721	1.010	1.886	2.201	1.551	27.895	6.880	6.368	30.046	90.926	0.140
	UMich [45]	0.135	1.531	0.971	2.029	4.596	0.448	0.062	0.062	0.040	0.109	0.176	0.020
	HKU-Mars [21]	1.065	1.414	1.572	2.527	5.458	0.779	0.165	0.075	0.122	0.246	0.893	0.036
	DVL [24]	0.261	**0.194**	**0.129**	0.391	0.879	**0.193**	0.022	0.016	**0.007**	0.030	0.078	0.013
	MIAS-LCEC (Ours)	**0.144**	0.234	0.175	**0.363**	**0.724**	0.225	**0.016**	**0.011**	0.011	**0.024**	**0.045**	**0.012**
Challenging Illumination	CRLF [27]	0.088	1.778	1.156	1.876	2.141	1.613	18.080	2.954	3.886	19.047	34.444	0.132
	UMich [45]	0.361	4.841	2.924	5.012	12.087	0.314	0.108	0.223	0.148	0.330	0.679	0.043
	HKU-Mars [21]	7.762	8.774	7.235	14.996	38.750	0.338	1.609	1.578	1.679	3.386	10.520	0.034
	DVL [24]	0.546	1.573	0.256	1.747	8.466	0.207	0.328	0.144	0.099	0.377	1.970	**0.022**
	MIAS-LCEC (Ours)	**0.347**	**0.524**	**0.180**	**0.749**	**1.554**	**0.149**	**0.087**	**0.057**	**0.030**	**0.118**	**0.220**	0.033
All	CRLF [27]	0.197	1.625	0.946	1.683	2.201	1.465	10.051	3.036	2.589	11.133	140.316	0.132
	UMich [45]	0.485	1.945	4.272	4.265	22.778	0.214	0.217	0.134	0.115	0.333	2.394	0.020
	HKU-Mars [21]	2.140	2.156	1.988	3.941	38.750	0.118	0.806	0.555	0.475	1.261	10.520	0.034
	DVL [24]	0.201	0.292	**0.104**	0.423	8.466	0.037	0.075	0.050	0.026	0.100	1.970	0.013
	MIAS-LCEC (Ours)	**0.133**	**0.196**	0.110	**0.298**	**1.554**	**0.012**	**0.040**	**0.028**	**0.019**	**0.061**	**0.220**	**0.012**

D. Comparisons with State-of-the-Art Methods

Quantitative comparisons with SoTA approaches on the MIAS-LCEC-TF70 and MIAS-LCEC-TF360 datasets are presented in Tables I and II. Additionally, qualitative results for these datasets are illustrated in Figs. 7, 8, and 9. It is important to note that the results from the first iteration of MIAS-LCEC are reported here because the accuracy achieved is already higher than that of existing SoTA approaches.

The results shown in Table I suggest that our method outperforms all other SoTA approaches on a total of 60 scenarios, all captured using a Livox Mid-70 LiDAR. Specifically, MIAS-LCEC reduces e_r by around 30-93% and decreases e_t by 39-99%, compared to existing SoTA algorithms. We attribute these performance improvements to the coarse-to-fine correspondence matching pipeline based on LVMs, which sets strict criteria for reliable sparse correspondence selection and propagates these matches to generate dense correspondences, thereby improving the quality of the PnP solutions. It can also be observed that MIAS-LCEC achieves lower mean e_r and e_t values than all other approaches across the total six subsets. Our method dramatically outperforms CRLF, UMich, and HKU-Mars and is slightly better than DVL in scenarios with low noise and abundant features, while it performs significantly better than all methods in challenging conditions, particularly under poor illumination and adverse weather, or when few geometric features are detectable. This impressive performance can be attributed to MobileSAM, a powerful LVM, capable of learning informative, general-purpose deep

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3456299

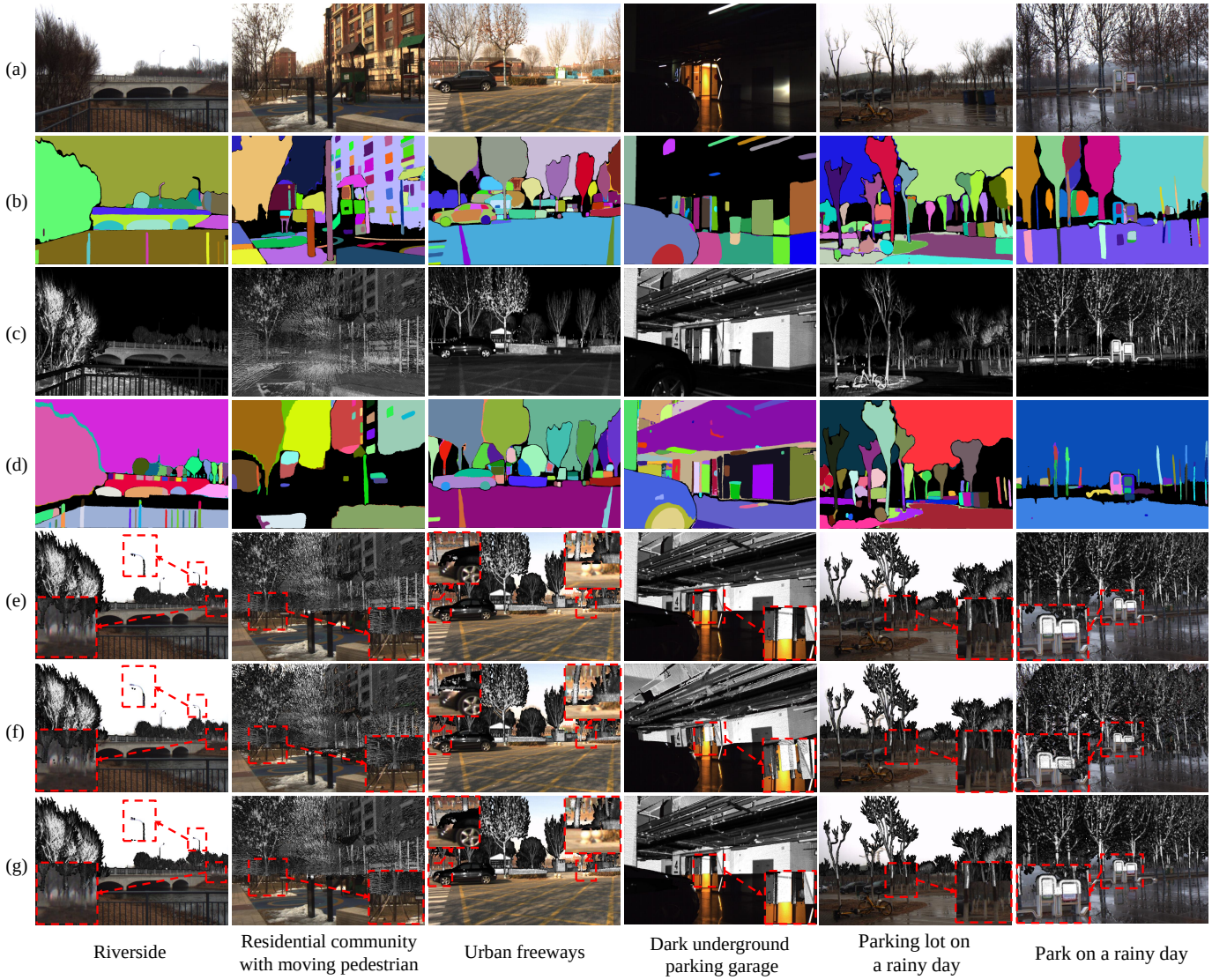


Fig. 7. Qualitative comparisons with SoTA target-free LCEC approaches on the MIAS-LCEC-TF70 dataset: (a)-(b) RGB images and their segmentation results; (c)-(d) LIP images and their segmentation results; (e)-(g) experimental results achieved using MIAS-LCEC (ours), HKU-Mars, and DVL, shown by merging LIP and RGB images, where significantly improved regions are shown with red dashed boxes.

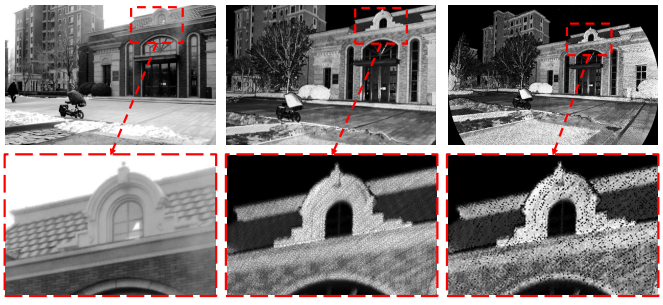


Fig. 8. Qualitative comparison between our proposed MIAS-LCEC and DVL in terms of LIP image generation on the MIAS-LCEC-TF70 dataset.

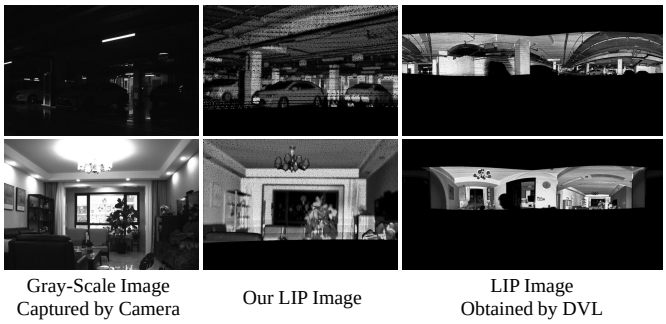


Fig. 9. Qualitative comparison between our proposed MIAS-LCEC and DVL in terms of LIP image generation on the MIAS-LCEC-TF360 dataset.

features for robust image segmentation. Surprisingly, as observed in Fig. 7, MobileSAM can effectively segment both RGB and LIP images captured in challenging conditions, such as dark underground parking garages or during rainy days,

TABLE II
QUANTITATIVE COMPARISON OF OUR PROPOSED MIAS-LCEC APPROACH WITH OTHER SOTA ONLINE, TARGET-FREE APPROACHES ON THE MIAS-LCEC-TF360 DATASET, WHERE THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

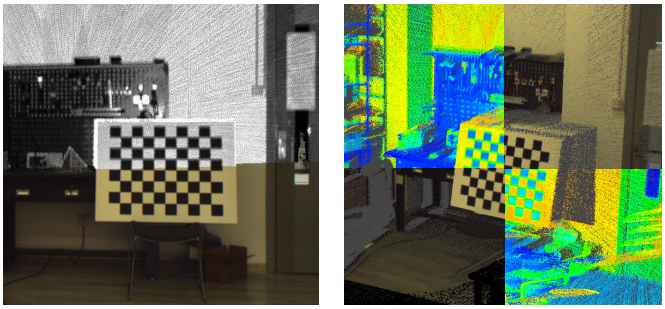
Error	Approach	Indoor	Outdoor
e_r (°)	CRLF [27]	1.469	1.402
	UMich [45]	1.802	2.698
	HKU-Mars [21]	96.955	25.611
	DVL [24]	63.003	46.623
	MIAS-LCEC (Ours)	**0.963**	**0.659**
e_t (m)	CRLF [27]	13.484	0.139
	UMich [45]	0.200	0.135
	HKU-Mars [21]	4.382	9.914
	DVL [24]	0.919	1.778
	MIAS-LCEC (Ours)	**0.182**	**0.114**

where the objects are even unrecognizable to human observers.

Additionally, the results on the MIAS-LCEC-TF360 dataset somewhat exceed our expectations. From Table II, it is evident that while the other approaches achieve poor performance on this dataset, MIAS-LCEC demonstrates acceptable performance, indicating strong adaptability to more challenging scenarios, with narrow overlapping areas between LIP and RGB images. This performance improvement is primarily attributed to our developed LIP image generation strategy, which incorporates several image pre-processing techniques in our practical implementation to refine the LIP images and align the FoVs between the two sensors as closely as possible. As illustrated in Figs. 8 and 9, the LIP image generated by DVL contains numerous holes and has a significantly different FoV compared to the RGB image, resulting in unexpected false correspondence matches, which can deteriorate the algorithm's efficiency and robustness. In contrast, MIAS-LCEC can generate LIP images that look as if taken from the same perspective to the actual camera, thus improving the performance of cross-modal mask matching.

E. Comparison with An Offline, Target-Based Approach

This subsection presents additional experimental results on the MIAS-LCEC-CB70 dataset, comparing our approach with ACSC [35], a SoTA offline, target-based LCEC algorithm, when calibration targets are available. The checkerboard corner points in both the LIP and RGB images are used as ground-truth correspondences. In contrast to ACSC, which employs ground-truth correspondences to estimate the extrinsic parameters, our approach conducts online, target-free LCEC. The reprojection errors of these correspondences are computed to quantify the performance of both algorithms. As illustrated in Fig. 10, the visualization of LCEC calibration results through LiDAR and camera data fusion suggests the high accuracy of our approach. Furthermore, as shown in Table III, while MIAS-LCEC achieves satisfactory results, its performance is slightly inferior to that of ACSC. This observation is within our expectations, as ACSC directly minimizes the mean reprojection error of the ground-truth correspondences to determine extrinsic parameters. In contrast, our method relies on distinguishable and matchable masks present in both modalities.



(a) (b)

Fig. 10. Visualization of LCEC calibration results through LiDAR and camera data fusion: (a) fusion of LIP and RGB images; (b) LiDAR point cloud partially rendered by color.

TABLE III
COMPARISON OF REPROJECTION ERRORS BETWEEN OFFLINE CALIBRATION AND OUR PROPOSED MIAS-LCEC APPROACH ON THE MIAS-LCEC-CB70 DATASET.

Yaw Angle	Approach	Distance from Checkerboard			Average ϵ
		3 m	4 m	5 m	(pixel)
$+30°$	Offline [35]	1.451	1.359	1.307	1.372
	MIAS-LCEC (Ours)	1.445	1.626	1.653	1.575
$+15°$	Offline [35]	1.589	1.544	1.361	1.498
	MIAS-LCEC (Ours)	1.919	1.473	1.386	1.593
$0°$	Offline [35]	1.664	2.132	1.329	1.708
	MIAS-LCEC (Ours)	1.593	2.086	1.608	1.762
$-15°$	Offline [35]	1.539	1.838	1.743	1.706
	MIAS-LCEC (Ours)	1.572	1.802	2.204	1.859
$-30°$	Offline [35]	1.439	1.534	1.470	1.481
	MIAS-LCEC (Ours)	1.569	2.020	1.883	1.824

F. LCEC Performance with Increasing Iterations

Fig. 11 illustrates the accuracy of our algorithm with respect to an increasing number of iterations. It is evident that (1) after the first iteration, our approach attains satisfactory accuracy, and (2) after the third iteration, its performance stabilizes and remains relatively consistent (the values of e_r and e_t decrease by approximately 26% and 11%, respectively, from the first to the sixth iterations). Therefore, we contend that a single iteration suffices for our MIAS-LCEC approach, striking a balance between accuracy and efficiency. However, additional iterations can certainly be considered when computational resources are abundant.

V. CONCLUSION

This article introduced MIAS-LCEC, a fully online, target-free LiDAR-camera extrinsic calibration approach, developed based on a state-of-the-art large vision model. Compared to prior arts, our approach is more capable of matching cross-modal features and outperforms existing state-of-the-art algorithms. To benefit the robotic community, we also designed a

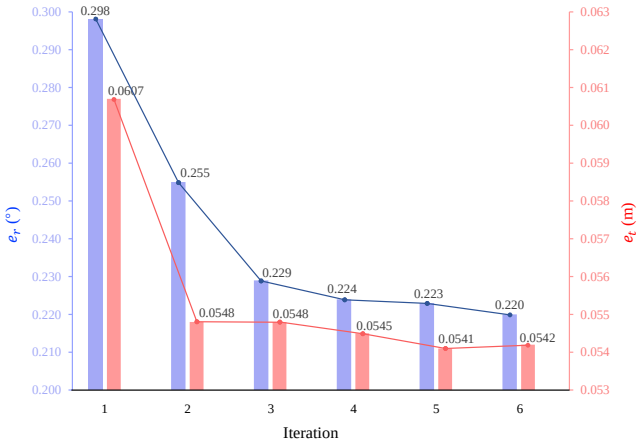


Fig. 11. Performance of MIAS-LCEC with increasing iterations on the MIAS-LCEC-TF70 dataset.

calibration toolbox with an interactive visualization interface based on our developed approach. Extensive experiments were conducted on three real-world datasets to comprehensively evaluate the performance of MIAS-LCEC. The experimental results demonstrate that (1) MIAS-LCEC achieves robust and accurate LiDAR-camera extrinsic calibration without the need for any targets, (2) it demonstrates high adaptability to diverse challenging scenarios by introducing a virtual camera with iterative pose updates to generate more accurate LiDAR intensity projections, and (3) the SoTA image segmentation LVM is successfully applied for this specific task by detecting distinguishable and matchable masks across different modalities. While achieving high accuracy and robustness, the real-time performance of our algorithm still requires improvement, a task we will address in future work.

REFERENCES

[1] E. Arnold *et al.*, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[2] R. Fan *et al.*, *Autonomous driving perception*. Springer, 2023.

[3] X. Bai *et al.*, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1090–1099.

[4] Y. Ai *et al.*, "LiDAR-camera fusion in perspective view for 3D object detection in surface mine," *IEEE Transactions on Intelligent Vehicles*, 2023, DOI: 10.1109/TIV.2023.3343377.

[5] R. Fan *et al.*, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356.

[6] Y. Cui *et al.*, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.

[7] R. Fan *et al.*, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[8] Y. Li *et al.*, "DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 182–17 191.

[9] J. Li *et al.*, "RoadFormer: Duplex transformer for rgb-normal semantic road scene parsing," *IEEE Transactions on Intelligent Vehicles*, 2024, DOI: 10.1109/TIV.2024.3388726.

[10] L. Zhang *et al.*, "FS-Net: LiDAR-camera fusion with matched scale for 3D object detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2023, DOI: 10.1109/TITS.2023.3287557.

[11] J. Huang *et al.*, "RoadFormer+: Delivering rgb-x scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Transactions on Intelligent Vehicles*, 2024, DOI: 10.1109/TIV.2024.3448251.

[12] R. Fan *et al.*, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2019.

[13] Y. Zhu *et al.*, "Online camera-LiDAR calibration with sensor semantic information," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4970–4976.

[14] Z. Wu *et al.*, "S^{3}M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 2, pp. 3940–3951, 2024.

[15] Y. Sun *et al.*, "ATOP: An attention-to-optimization approach for automatic LiDAR-camera calibration via cross-modal object matching," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 696–708, 2022.

[16] N. Ou *et al.*, "Targetless LiDAR-camera calibration via cross-modality structure consistency," *IEEE Transactions on Intelligent Vehicles*, 2023, DOI: 10.1109/TIV.2023.3337490.

[17] H. Zhao *et al.*, "Dive deeper into rectifying homography for stereo camera online self-calibration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 479–14 485.

[18] S. Tang *et al.*, "Robust calibration of vehicle solid-state LiDAR-camera perception system using line-weighted correspondences in natural environments," *IEEE Transactions on Intelligent Transportation Systems*, 2023, DOI: 10.1109/TITS.2023.3328062.

[19] F. Lv and K. Ren, "Automatic registration of airborne LiDAR point cloud data and optical imagery depth map based on line and points features," *Infrared Physics & Technology*, vol. 71, pp. 457–463, 2015.

[20] S. Wang *et al.*, "Temporal and spatial online integrated calibration for camera and LiDAR," in *2022 IEEE International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 3016–3022.

[21] C. Yuan *et al.*, "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.

[22] J. Castorena *et al.*, "Autocalibration of LiDAR and optical cameras via edge alignment," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2862–2866.

[23] X. Zhang *et al.*, "Line-based automatic extrinsic calibration of LiDAR and camera," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9347–9353.

[24] K. Koide *et al.*, "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.

[25] C. Ye *et al.*, "Keypoint-based LiDAR-camera online calibration with robust geometric network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2021.

[26] Y. Wang *et al.*, "Automatic registration of point cloud and panoramic images in urban scenes based on pole matching," *International Journal of Applied Earth Observation and Geoinformation*, vol. 115, p. 103083, 2022.

[27] T. Ma *et al.*, "CRLF: Automatic calibration and refinement based on line feature for LiDAR and camera in road scenes," *arXiv preprint arXiv:2103.04558*, 2021.

[28] J. Li *et al.*, "Automatic registration of panoramic image sequence and mobile laser scanning data using semantic features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 136, pp. 41–57, 2018.

[29] Y. Liao *et al.*, "SE-Calib: Semantic edges based LiDAR-camera boresight online calibration in urban scenes," *IEEE Transactions on Geoscience and Remote Sensing*, 2023, DOI: 10.1109/TGRS.2023.3278024.

[30] Y. Han *et al.*, "Auto-calibration method using stop signs for urban autonomous driving applications," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 179–13 185.

[31] Kirillov *et al.*, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023, pp. 4015–4026.

[32] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2023.

[33] C.-W. Liu *et al.*, "Playing to vision foundation model's strengths in stereo matching," *IEEE Transactions on Intelligent Vehicles*, 2024.

[34] C. Zhang *et al.*, "Faster Segment Anything: Towards lightweight sam for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.

[35] J. Cui *et al.*, "ACSC: Automatic calibration for non-repetitive scanning solid-state LiDAR and camera systems," *arXiv preprint arXiv:2011.08516*, 2020.

[36] J. Beltrán *et al.*, "Automatic extrinsic calibration method for LiDAR and camera sensor setups," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 677–17 689, 2022.

[37] G. Koo *et al.*, "Analytic plane covariances construction for precise planarity-based extrinsic calibration of camera and LiDAR," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6042–6048.

[38] Y. Xie *et al.*, "A4LidarTag: Depth-based fiducial marker for extrinsic calibration of solid-state LiDAR and camera," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6487–6494, 2022.

[39] G. Yan *et al.*, "Joint camera intrinsic and LiDAR-camera extrinsic calibration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 446–11 452.

[40] D. Tsai *et al.*, "Optimising the selection of samples for robust LiDAR camera calibration," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2631–2638.

[41] D. Scaramuzza *et al.*, "Extrinsic self calibration of a camera and a 3D laser range finder from natural scenes," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2007, pp. 4164–4169.

[42] A. Dhall *et al.*, "LiDAR-camera calibration using 3D-3D point correspondences. arxiv 2017," *arXiv preprint arXiv:1705.09785*.

[43] S. Bileschi, "Fully automatic calibration of LiDAR and video streams from a vehicle," in *2009 IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2009, pp. 1457–1464.

[44] G. Pandey *et al.*, "Automatic targetless extrinsic calibration of a 3D LiDAR and camera by maximizing mutual information," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 26, no. 1, 2012, pp. 2053–2059.

[45] G. Pandey *et al.*, "Automatic extrinsic calibration of vision and LiDAR by maximizing mutual information," *Journal of Field Robotics*, vol. 32, no. 5, pp. 696–722, 2015.

[46] M. Pu *et al.*, "EDTER: Edge detection with Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1402–1412.

[47] N. Schneider *et al.*, "RegNet: Multimodal sensor registration using deep neural networks," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1803–1810.

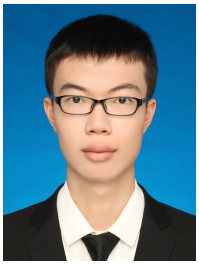
[48] G. Iyer *et al.*, "CalibNet: Geometrically supervised extrinsic calibration using 3D spatial transformer networks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1110–1117.

[49] X. Lv *et al.*, "LCCNet: LiDAR and camera self-calibration using cost volume network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2894–2901.

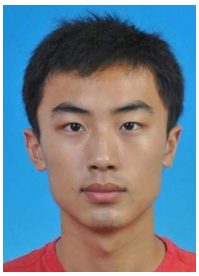
[50] A. Geiger *et al.*, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[51] P.-E. Sarlin *et al.*, "SuperGlue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4938–4947.

[52] B. Zhang and R. T. Rajan, "Multi-FEAT: Multi-feature edge alignment for targetless Camera-LiDAR calibration," *arXiv preprint arXiv:2207.07228*, 2022.



Zhiwei Huang (Graduate Student Member, IEEE) received the B.E. degree in automation from Tongji University, Shanghai, China, in 2024. He is currently pursuing his M.Sc. degree, supervised by Prof. Rui Fan, with the MIAS Group in the College of Electronics and Information Engineering at Tongji University. His research interests include computer vision and robotics.



Yikang Zhang (Graduate Student Member, IEEE) obtained his B.Sc. degree in Automotion from the Beijing Institute of Technology in 2017, followed by an M.S. degree in ECE from UMASS Amherst in 2019, where he specialized in Model Predictive Control and Physical Unclonable Functions. He has since gained extensive industry experience, working with companies such as Tusimple, where he focused on truck state machine and controller, UnityDrive, where he contributed to Path Planner and chassis control, and CASIA, where he worked on Trajectory Prediction Algorithms. Currently, Yikang is pursuing his Ph.D. degree, supervised by Prof. Rui Fan. His research interests include Simulation, Planning, and Control in complex environment.

Qijun Chen (Senior Member, IEEE) received the B.S. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999. He is currently a Full Professor in the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include robotics control, environmental perception, and understanding of mobile robots and bioinspired control.

Rui Fan (Senior Member, IEEE) received the B.Eng. degree in Automation from the Harbin Institute of Technology in 2015 and the Ph.D. degree (supervisors: Prof. John G. Rarity and Prof. Naim Dahnoun) in Electrical and Electronic Engineering from the University of Bristol in 2018. He worked as a Research Associate (supervisor: Prof. Ming Liu) at the Hong Kong University of Science and Technology from 2018 to 2020 and a Postdoctoral Scholar-Employee (supervisors: Prof. Linda M. Zangwill and Prof. David J. Kriegman) at the University of California San Diego between 2020 and 2021. He began his faculty career as a Full Research Professor with the College of Electronics & Information Engineering at Tongji University in 2021, and was then promoted to a Full Professor in the same college, as well as at the Shanghai Research Institute for Intelligent Autonomous Systems in 2022.

Prof. Fan served as an associate editor for ICRA'23 and IROS'23/24, an area chair for ICIP'24, and a senior program committee member for AAAI'23/24/25. He is the general chair of the AVVision community and organized several impactful workshops and special sessions in conjunction with WACV'21, ICIP'21/22/23, ICCV'21, and ECCV'22. He was honored by being included in the Stanford University List of Top 2% Scientists Worldwide in both 2022 and 2023, recognized on the Forbes China List of 100 Outstanding Overseas Returnees in 2023, and acknowledged as one of Xiaomi Young Talents in 2023. His research interests include computer vision, deep learning, and robotics, with a specific focus on humanoid visual perception under the two-streams hypothesis.