SE-Calib: Semantic Edge-Based LiDAR–Camera Boresight Online Calibration in Urban Scenes

Youqi Liao, Jianping Li, *Member, IEEE*, Shuhao Kang, Qiang Li, Guifang Zhu, Shenghai Yuan, Zhen Dong, *Member, IEEE*, and Bisheng Yang

***Abstract*— Rigorous boresight calibration between light detection and ranging (LiDAR) and the camera is crucial for geometry and optical information fusion in Earth observation and robotic applications. Although boresight parameters can be obtained through precalibration with artificial targets, unforeseen movement of sensors during data collection can lead to significant errors in the boresight parameters. To address this issue, we propose SE-Calib, an automatic and target-free online boresight calibration method for LiDAR–camera systems. SE-Calib first extracts the semantic edge features from both point clouds and images simultaneously using the 3-D semantic segmentation (3-D-SS) and 2-D semantic edge detection (2-D-SED) methods. The boresight parameters are then optimized with an adaptive solver and maximizing the soft semantic response consistency metric (SSRCM) scores iteratively. The SSRCM is designed to evaluate the coherence of cross-modular semantic edge features, and a confidence function is proposed to filter out unreliable optimization results. Experiments conducted on challenging urban datasets show an average boresight error of 0.206° (2.47 pixels in reprojection error), demonstrating the effectiveness and robustness of the proposed method.**

***Index Terms*— Boresight parameters, mobile mapping system (MMS), semantic edge features, sensors calibration.**

I. Introduction

THE fusion of point clouds and images is crucial for various applications in Earth observation and robotics, such as the mobile mapping system (MMS) [1], [2], simultaneous localization and mapping (SLAM) [3], [4], [5], 3-D semantic segmentation (3-D-SS) [6], [7], [8], and 3-D object detection (3-D-OD) [9], [10]. While the camera provides dense color information and texture, images are ambiguous in depth sensing and unreliable under harsh illumination conditions. On the other hand, light detection and ranging (LiDAR) offers accurate and light-invariant depth information but lacks density and texture [6]. Combining sensors of multiple modalities provides complementary environmental information and reduces data uncertainty.

Determining the extrinsic transform relationship (boresight and translation parameters) between LiDAR and the camera is an essential step in integrating the 2-D and 3-D information [11]. While precalibration with artificial targets is commonly performed in a calibration field [12], [13], the accuracy of these precalibrated parameters decreases over time due to sensor movement during data collection [14]. In particular, large errors in rotation parameters are more common on vehicles than errors in translation parameters [15]. Therefore, an accurate and reliable target-free online boresight calibration method is an urgent requirement for effectively fusing geometry and optical information. In recent years, various target-free methods have been developed for aligning specific features extracted from point clouds and images, e.g., line features [16], [17], edge features [18], [19], and semantic features [14], [20], [21]. However, matching of stable correspondences between noisy line features or edge features can be challenging in real-world scenes, leading to solvers easily getting stuck in locally optimal values [11], [22]. Most existing semantic feature-based methods rely on specific semantic classes and are limited to specific scenes, such as parking vehicles [21], poles [14], and stop signs [23], which may not be available in many frames. Moreover, a crucial issue, the reliability of the boresight parameter estimation, is often ignored in these methods. Ultimately, the accuracy, robustness, and generalization of existing methods are still limited. Thus, our motivation is to find one kind of primitive that is accurate, robust, and generative enough for the online boresight calibration task in the urban scene.

Due to the wide distribution of semantic objects and the distinction of edge features, semantic edge inherits the above excellent characteristics and suppresses the matching outliers. Therefore, we propose the SE-Calib to automatically estimate the boresight (extrinsic rotation) parameters between a laser scanner and a monocular camera by aligning the 2-D–3-D

Manuscript received 23 February 2023; revised 24 April 2023 and 13 May 2023; accepted 16 May 2023. Date of publication 19 May 2023; date of current version 1 June 2023. This work was supported in part by the National Natural Science Foundation Project under Grant 42130105 and Grant 42201477; in part by the China Postdoctoral Science Foundation under Grant 2022M712441 and Grant 2022TQ0234; and in part by the Open Fund of Key Laboratory of Urban Spatial Information, Ministry of Natural Resources, under Grant 2023ZD001. *(Corresponding author: Jianping Li.)*

Youqi Liao and Zhen Dong are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) and the Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China (e-mail: martin_liao@whu.edu.cn; dongzhenwhu@whu.edu.cn).

Jianping Li is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, and also with the Key Laboratory of Urban Spatial Information, Ministry of Natural Resources, Beijing 100044, China (e-mail: jianping.li@ntu.edu.sg).

Shuhao Kang and Bisheng Yang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430079, China (e-mail: shkang@whu.edu.cn; bshyang@whu.edu.cn).

Qiang Li and Guifang Zhu are with Shenyang Geotechnical Investigation and Surveying Research Institute Company Ltd., Shenyang 110004, China (e-mail: q.lee@163.com; gfzhu200909@163.com).

Shenghai Yuan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: shyuan@ntu.edu.sg).

Digital Object Identifier 10.1109/TGRS.2023.3278024

semantic edge primitives. The proposed method contains two main steps: semantic edge features extraction and adaptive boresight optimization. Specifically, the 2-D semantic edge features extracted from images are achieved by an end-to-end deep-learning-based 2-D semantic edge detection (2-D-SED) method [24]. With respect to the point clouds, the 3-D semantic edge features are extracted with the 3-D-SS method [8] and a lightweight semantic edge detection algorithm. Then, using the soft semantic response consistency metric (SSRCM), adaptive boresight optimization evaluates the responsive degrees and aligns extracted semantic edge features iteratively. Then, the confidence function is utilized as a plug-and-play module to remove unreliable calibration results. The main contributions are given as follows.

1) The consistency between 2-D and 3-D semantic edge features from different modalities is leveraged for LiDAR–camera boresight calibration. The SE-Calib benefits from the robustness and widespread availability of semantic edge features, resulting in both accurate and robust calibration results in various urban environments.
2) The SSRCM is proposed to weight semantic edge correspondence pairs based on their response level and to filter out unreliable boresight estimation results. It fully leverages semantic edge features in the optimization process to achieve the global optimal solution.

The rest of this article is organized as follows. Section II provides a review of the related works. Section III details the proposed method. Comprehensive experiments to evaluate the SE-Calib are conducted in Section IV, followed by a discussion of key components, analysis of features extraction noise, hyperparameters settings, and false cases in Section V. Finally, conclusions are drawn in Section VI.

II. Related Work

In recent decades, there have been numerous reports of LiDAR–camera calibration works in the photogrammetry and robotics fields. In this section, we broadly divide existing investigations into two categories: target-based methods and target-free methods. In light of the recent expansion of deep-learning-based research in multimodal registration and localization, relevant studies are also covered.

A. Target-Based Methods

Early research in the area mainly utilized artificial targets as calibration constraints. Zhang and Pless [25] introduced an algorithm for calibrating a camera and 2-D laser range finder using a planar checkerboard. They extracted checkpoints from laser readings and detected the corresponding checkboard grids on the image. Then, the rotation and translation parameters are calculated using a linear solution. Unnikrishnan and Hebert [26] expanded Zhang's method [25] by manually selecting keypoints from both images and point clouds and made their MATLAB toolbox publicly available. Later, Scaramuzza et al. [27] proposed to register the LiDAR and camera based on the perspective-n-points (PnP) algorithm with manually selected feature points and pixels. The above methods belong to point-to-plane methods. However, the drawbacks of manually selecting point-pixel pairs, such as artificial error and time-consuming, have led to the development of various object-based methods, including sphere [13], [28], circular [29], trihedron [30], and triangular checkerboard [31]. Despite these advancements, all target-based methods still rely on specific artificial targets and are unsuitable for challenging urban scenes.

B. Target-Free Methods

In the last decade, there has been a surge in the development of target-free methods for LiDAR and camera calibration. We divide them into five streams: traditional feature-based methods, mutual information (MI)-based methods, semantic feature-based methods, motion-based methods, and end-to-end learning-based methods.

1) Traditional Feature-Based Methods: In our knowledge, Levinson and Thrun [15] proposed the first target-less calibration research based on edge features alignment. They extracted geometric edges from point clouds and spectrum edges [32] from images as candidate features for registration and maximized the global response value to refine the extrinsic transform parameters. Taylor et al. [33] improved the method in [15] by considering the edge orientation and using the gradient orientation measure (GOM) [34] for LiDAR and camera registration. Castorena et al. [18] performed data fusion first and then used the alignment of depth and intensity edges to correct the transform parameters. Zhang et al. [22] proposed a line-based method that registered the line features extracted from images and point clouds. Wang et al. [17] further improved the line-based method in [22] by considering temporal synchronization. However, these traditional target-free methods have some limitations. The geometric edge/line features and spectrum edge/line features extracted from point clouds and images are based on different rules and not strongly correlated, which is exacerbated under changing illumination.

2) MI-Based Method: MI has also been used as a specific feature for LiDAR–camera calibration by exploiting the interior relevance between the intensity of point clouds and grayscale of images [35], [36], [37]. However, the correlation between LiDAR and image may not be stable due to variability of devices and scanning modes [22].

3) Semantic Feature-Based Methods: With the recent development of deep learning, several studies [11], [14], [17], [20], [21], [23], [38] have focused on semantic features for calibration. Li et al. [21] detected parking vehicles first and registered point clouds and images with extracted vehicles afterward. Zhu et al. [11] maximized the overlapping area of vehicles from point clouds and images. Ma et al. [20] registered the LiDAR and camera by aligning road lanes and poles. Han et al. [23] extracted stop signs as primitives for calibration and temporally updated results with a Kalman filter. Jiang et al. [38] explored the consistency of semantic labels between point clouds and images. Wang et al. [14] presented an automatic registration method based on pole matching. Despite their impressive results, the generalization of these semantic-based methods is still restricted by three factors:

first, the reliance on specific semantic objects circumscribes the availability of challenging scenes; second, the blurry boundaries of semantic segmentation results can affect the accuracy of calibration; and third, the reliability of calibrated parameters is overlooked.

4) Motion-Based Methods: [39], [40], [41], and [42] estimated the extrinsic parameters of LiDAR and camera by utilizing the motion of sensors as auxiliary cues. Structure-from-motion (SfM) algorithms are used to solve visual odometry (VO), and the iterative closest point (ICP) algorithm is used to solve LiDAR odometry. Taylor and Nieto [39] proposed a motion-based approach to automatically calibrate any number of cameras, 3-D LiDARs, and inertial measurement unit (IMU); then, Taylor and Nieto [40] further incorporated existing motion-based pipelines into a novel probabilistic model to calibrate mobile system's sensors. Corte et al. [41] proposed a unified framework to optimize the intrinsics, extrinsics, and time delay of several sensors. Sen et al. [42] proposed the SceneCalib that used all sensor measurements simultaneously to constrain all the parameters in a multicamera/single-LiDAR system, while the accuracies of these methods are highly dependent on the motion estimation.

5) End-to-End Learning-Based Methods: [43], [44], [45], and [46] utilized the neural networks to regress the miscalibration error of given extrinsic parameters directly. RegNet [43] first presented a convolutional neural network (CNN) to estimate the six-degree-of-freedom extrinsic parameters, while iterative estimation and temporal filtering were proposed to refine calibration results. As the loss function minimizes the distance between predicted extrinsic parameters and ground-truth ones, RegNet needs to be retrained when sensor intrinsic parameters change. CalibNet [44] improved the RegNet by maximizing the geometric and photometric consistency of the point clouds and images to regress the extrinsic parameters implicitly. LCCNet [45] proposed an online self-calibration method by introducing the cross-attention module to measure the similarity between point clouds and images. SST-Calib [46] explored the semantic consistency and VO information to regress extrinsic parameters. However, the dependence on specific sensor configurations (both LiDAR and camera) makes them hard to implement.

C. Relevant Deep-Learning Studies

Recently, there has been a growing interest in learning-based description and detection of local features [47] for image-to-image and image-to-point cloud registration. This approach tries to overcome the limitations of traditional methods by learning cross-domain representation simultaneously. For example, LCD [48] proposed a dual autoencoder network learning descriptors for cross-modality matching. 2D3D-MatchNet [49] extracted keypoints from image and point clouds using scale invariant feature fransform (SIFT) and intrinsic shape signatures (ISS) detectors, respectively; then, a triplet-like CNN is used to learn the descriptors for keypoints. However, the poor top-1 feature matching success rate, about 20%, indicates that the response degree of SIFT and ISS detectors is low. The large deviation of registration results (about 7° of average rotation error and 1.5 m of average translation error) indicates that artificial detectors for multiple modalities registration are still at an early stage. P2-Net [50] addressed this issue by jointly learning the detectors and descriptors of images and point clouds and then mapping 2-D and 3-D local features to a high-dimensional space for latent response extraction. However, due to the significant difference in modalities and sparsity of outdoor laser point clouds, P2-Net is only suitable for indoor registration tasks. In the area of camera relocation, DeepI2P [51] and CorrI2P [52] predicted the overlapping region between the frustum of the camera and LiDAR and then used an optimizer to find the solution. However, the newest and most accurate method, CorrI2P, only achieved the relative rotation error (RRE) $2.07° \pm 1.64°$ and $2.65° \pm 1.93°$ on KITTI [53] and Nuscenes [54] datasets, respectively, indicating the need for further improvement in accuracy.

Generally speaking, existing target-free calibration methods for LiDAR–camera are limited in terms of robustness, generalization, and reliability. Most early studies lack robustness because of concentrating on low-level artificial features. Recent studies that exploit semantic features are constrained to specific environments. Motion-based methods need to recover the trajectory of cameras and LiDARs. End-to-end learning-based methods are constrained by specific sensors configuration. In this article, we propose SE-Calib, an online calibration approach for the LiDAR and camera. With distinct and ubiquitous semantic edge features and tailored response assessment rules SSRCM, our method delivers accurate and robust boresight estimation in various urban environments.

III. Methodology

For a convenient description, point clouds and image captured by the laser scanner and camera are written as $\boldsymbol{P}$ and $\boldsymbol{I}$, respectively. For each laser point $\boldsymbol{P}_k$ and corresponding projection pixel $\boldsymbol{I}_{ij}$, the transformed relationship between the LiDAR and camera could be denoted as follows:

$$\begin{bmatrix} \boldsymbol{I}_{ij} \\ 1 \end{bmatrix} = \boldsymbol{K} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{P}_k \\ 1 \end{bmatrix} \tag{1}$$

where $\boldsymbol{R}$ and $\boldsymbol{t}$ are the boresight rotation matrix and translation vector from the laser scanner coordinates system to the camera coordinates system, respectively, and $\boldsymbol{K}$ is the precalibrated intrinsic matrix of the camera. Our work aims to estimate the boresight rotation matrix or rotation Euler angle vector $\boldsymbol{r}$ equally [as the 3-D rotation matrix lies in the Lie group $SO(3)$, $\boldsymbol{R} = \exp(\boldsymbol{r})$].

SE-Calib fully utilizes semantic edge features from point clouds and images and further leverages the correlation between cross-modality data to accurately estimate the boresight (extrinsic rotation) parameters between the camera and the laser scanner. The workflow of the proposed method is shown in Fig. 1. First, semantic edge features are extracted from images and point clouds. Then, an adaptive optimization process is performed based on the consistency between cross-modal semantic edge features. Moreover, a lightweight confidence check-up process (CCuP) is proposed to filter out unreliable frames. The key steps are detailed as follows.

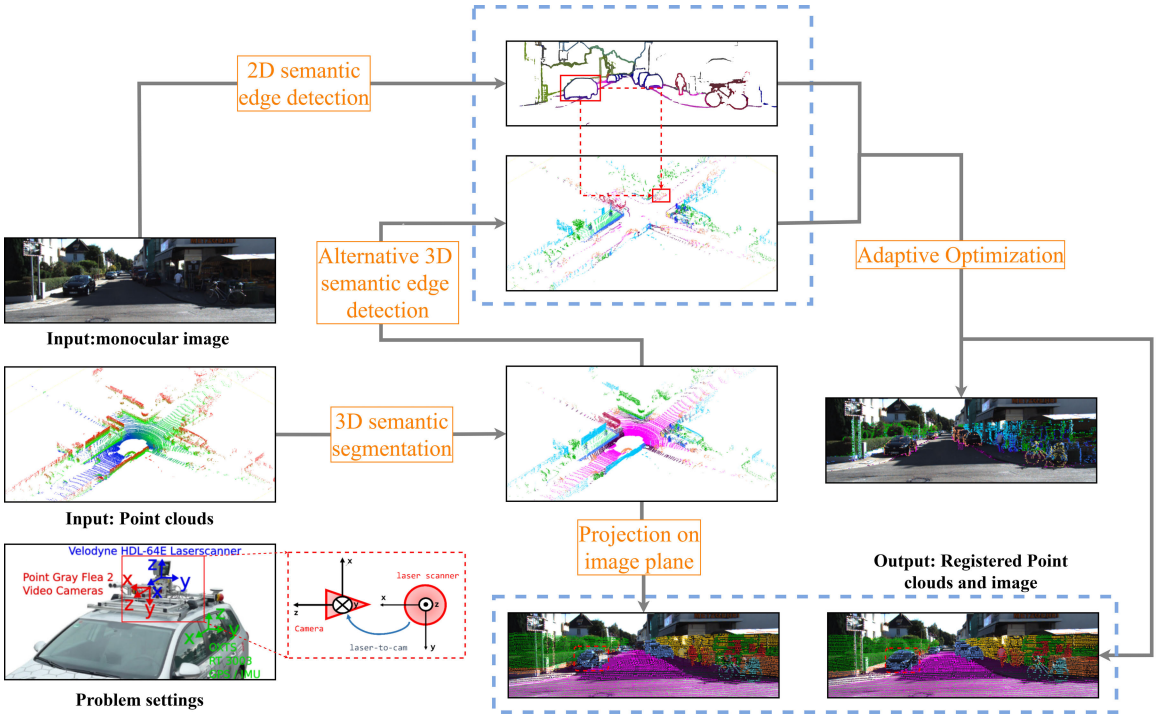

Fig. 1. Workflow of SE-Calib.

A. Semantic Edge Feature Extraction

1) 2-D Semantic Edge Extraction: Several CNN-based 2-D-SED methods [24], [55], [56] have been proposed to detect semantic edge features on images. In this work, we utilize the dynamic feature fusion (DFF) [24] network as our 2-D-SED backbone. The RGB image $\boldsymbol{I}$ is fed into the DFF network and a probability map set $\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_m\}$ obtained, where m indicates the number of categories, and each map has the same size as $\boldsymbol{I}$. After prediction, each pixel is labeled with a float vector (also referred to as soft target) indicating the probabilities of multiple semantic edge classes. Hinton et al. [57] elaborated that the soft target with higher entropy and lower variance compared to class labels (also referred to hard target) benefits the posterior aligning process implicitly. Furthermore, the hard target struggles to handle ambiguities caused by occlusion, where the contours of several foreground objects and background objects intersect at a single pixel (as shown in Fig. 2). Therefore, we utilize the soft target as calibration primitives to fully explore the uncertainty of the prediction. The advantages of the soft target will be discussed in Section V-B.

2) 3-D Semantic Edge Extraction: Despite being a growing field, there are a few approaches in 3-D semantic edge detection (3-D-SED), with previous studies primarily focusing on a single object or the indoor environment [58]. The 3-D-SED task requires both sophisticated local features and hierarchical high-level descriptors, which makes it difficult to transfer existing methods to outdoor sparse point clouds. Given the difficulties, we propose an alternative 3-D-SED method based on the mature 3-D-SS method. First, we feed point clouds into the 3-D-SS network [8], afterward extracting the probability map set (soft targets) and label map set (hard targets). Then, split point clouds into scans and detect label slips in two nearest neighbor points of each scan. If the current point has a different label from the prior or posterior one, it will be classified as an edge point. Finally, affix soft targets to detected edge points as semantic edge features. For details, the pseudocode of the proposed 3-D-SED method is shown in Algorithm 1. It should be retained that we are inspired by the similar geometric discontinuity strategy for edge detection in pioneering work [15].

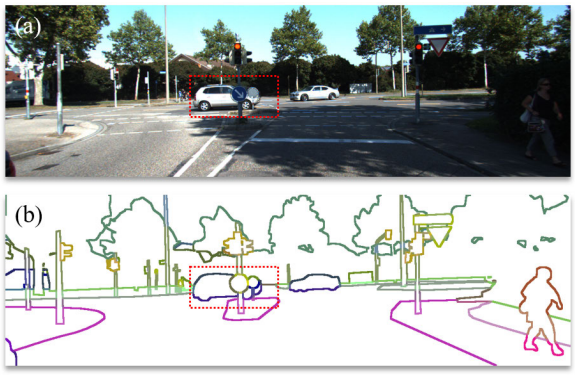

Fig. 2. Illustration of semantic edge ambiguities due to occlusion. (a) Monocular image and (b) semantic edge mask. Contours of background cars and foreground traffic sign intersect in the red block.

Compared with the widely investigated edge detection method based on distance discontinuities [5], [15] or local gradient [59], the proposed method is lightweight and aware of semantic boundaries without obvious distance change in scanning direction, e.g., boundaries of sidewalk and road. Fig. 3 visualizes the fully alternative 3-D-SED process.

Algorithm 1 Semantic Edge Detection on Point Cloud

Input: point clouds $\boldsymbol{P}$, semantic categories m, scans set $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$;
Initialize: semantic edge points set $\boldsymbol{P}_{se}$=∅;
3D-SS: $\boldsymbol{P} \xrightarrow{CNN}$ soft targets (semantic probabilities map) $\boldsymbol{F} \xrightarrow{solidfy}$ hard targets (label map) $\boldsymbol{L}$;
Semantic discontinuities search: split raw point clouds and hard targets into each scan: $\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_n\}$, $\boldsymbol{L} = \{\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_n\}$;
for i=1 to n: **do**
  for $j = 1$ to $|L_i| - 1$: **do**
    if $L_{S_i}^{j} \neq L_{S_i}^{j-1}$ or $L_{S_i}^{j} \neq L_{S_i}^{j+1}$ **then**
      pull $P_{S_i}^{j}$ into $\boldsymbol{P}_{se}$;
    end if;
  end for;
end for;
pick-up corresponding soft targets subset $\boldsymbol{F}_{se}$ from $\boldsymbol{F}$ with $\boldsymbol{P}_{se}$;
return $\boldsymbol{F}_{se}$, $\boldsymbol{P}_{se}$;

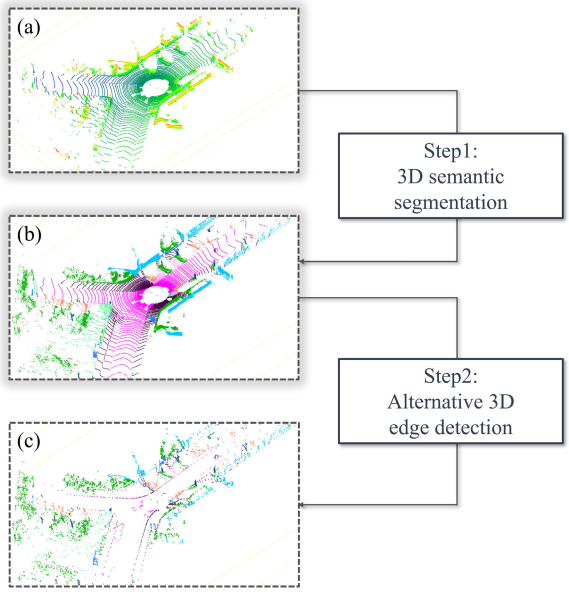

Fig. 3. Semantic edge detection process in point clouds. (a) Original point clouds. (b) 3-D-SS results. (c) Detected semantic edge points with the alternative 3-D-SED process. Rendering of (a) is according to the height value, while (b) and (c) are with the SemanticKITTI [60] colorization rules.

B. Adaptive Boresight Optimization

The extracted semantic edge features are used to estimate boresight parameters in our approach. Unlike prior works [61], [62], our method takes both accuracy and reliability into account, which is crucial for integrating the proposed method into existing perception systems.

The boresight optimization is based on the assumption that the semantic edge features from point clouds and images should be well-aligned if the parameters are unbiased. Based on the assumption, we project the semantic edge feature f_k from point clouds onto the image plane, find the corresponding semantic edge feature f_{ij}, and then assess the global semantic consistency with the SSRCM metric in the following

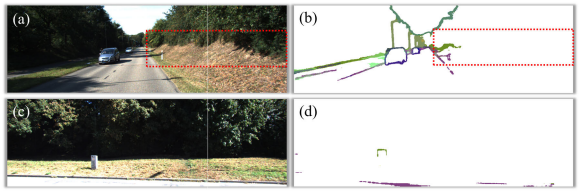

Fig. 4. Unreliable frames for calibration. The first column shows the origin RGB images and the second column shows the corresponding 2-D-SED prediction results. Red boxes in (a) and (b) false detection area, and (c) and (d) frames lack of meaningful features.

equation:

$$s(\boldsymbol{R}) = \sum_{f_k \in \boldsymbol{F}_{se}}^{f_k} f_{ij} \circ \boldsymbol{K}(\boldsymbol{R} f_k + \boldsymbol{t}). \tag{2}$$

The SSRCM scores point-pixel corresponding pairs adaptively and accumulate the scores among semantic edge features set to evaluate holistic response degree. In the adaptive optimization process, we solve the boresight parameters by maximizing the SSRCM function with a grid search method [15], [17], [22]. Specifically, centering around given initial boresight parameters with search step ssp, we perform the grid search across three dimensions of boresight parameters and compute $3^3 = 27$ different values of SSRCM scores. Then, we update the initial boresight parameters to the location with the highest score. A coarse-to-fine searching strategy is adopted for effectiveness and accuracy while avoiding local minimal:

1) beginning with a relatively large search step ssp to approach the vicinity of global optimum quickly;
2) halving the search step ssp if no higher score in neighbor grids.

Steps 1) and 2) are performed iteratively to approach the global optimum until the step size is smaller than the threshold.

Furthermore, we investigated the factors that compromise the calibration results mostly with our empirical experience and two issues were found:

1) false feature extraction results, as shown in Fig. 4(a) and (b);
2) lack of significant semantic edge features, as shown in Fig. 4(c) and (d).

To evaluate the reliability of optimized results, a confidence function is proposed as (3) (for intuitive description, we reimplement t denoting time stamp and $\boldsymbol{t}$ denoting translation vector without ambiguity there)

$$c^t(\boldsymbol{R}) = \frac{1}{N} \sum_{f_k \in \boldsymbol{F}_{se}}^{f_k} f_{ij}^t \circ \boldsymbol{K}(\boldsymbol{R} f_k^t + \boldsymbol{t}) = \frac{1}{N} s(\boldsymbol{R})^t. \tag{3}$$

A hard threshold to classify reliable and unreliable frames is stiff and unsuitable for various scenarios. Inspired by [7] and [63], spatial and temporal smoothness are utilized for reliable frames selection instead.

1) *Spatial Smoothness:* With correct boresight parameters, a large partition of pixel-point semantic edge features pairs should be highly responsive and the confidence value should keep at a high level.
2) *Temporal Smoothness:* With correct boresight parameters, the confidence value among the temporally nearest

Algorithm 2 CCuP

Input: optimized boresight parameters $\hat{\boldsymbol{r}}$, timestamp t, spatial smoothness threshold ϵ_s, temporal smoothness threshold ϵ_t;
calculate confidence score $c^t(\boldsymbol{r})$ by Eq. (4);
check optimization reliability:
if $c^t(\boldsymbol{r}) > \epsilon_s$ and $min(\frac{c^t(\boldsymbol{r})}{c^{(t-1)}(\boldsymbol{r})}, \frac{c^t(\boldsymbol{r})}{c^{(t+1)}(\boldsymbol{r})}, \frac{c^{(t+1)}(\boldsymbol{r})}{c^t(\boldsymbol{r})}, \frac{c^{(t-1)}(\boldsymbol{r})}{c^t(\boldsymbol{r})}) > \epsilon_t$ **then**
    $\hat{\boldsymbol{r}}^t_{reliable} = True$;
else
    $\hat{\boldsymbol{r}}^t_{reliable} = False$;
end if;
return confidence label $\hat{\boldsymbol{r}}^t_{reliable}$;

prior and posterior frames should be steady, without soaring or plunging.

Based on the considerations above, we design the CCuP by calculating the confidence score for each frame and signing those frames with smoothness in both spatial and temporal dimensions as reliable frames. The complete pseudocode of the CCuP is shown in Algorithm 2.

IV. Experiments

In this section, extensive experiments are conducted to validate SE-Calib on the KITTI [53] dataset. We use thousands of frames from the KITTI Odometry Benchmark as our test ground and verify the boresight parameters calibration results. KITTI is a typical urban road scenario dataset, including a large number of vehicles, pedestrians, buildings, traffic signs, and vegetation.

A. Dataset and Implementation

1) Dataset: KITTI [53] is a driving-scenes dataset for several tasks, including object detection, depth prediction, and semantic segmentation. The KITTI Odometry Benchmark was collected in German with a Velodyne-HDLE64 LiDAR and two high-resolution RGB cameras (only images from the left camera are used in our validations). The laser scanner captured ten frames a second and cameras took images with the same frequency. Data used in our experiments are synced and rectified. The transformation matrix from the calibration file is thought of as the ground truth in our experiment.

2) Model Training: For the image sequences, we pretrained the 2-D-SED model DFF on the Cityscapes [64] dataset with fine annotation and utilized on the KITTI dataset directly. For the point cloud sequences, SemanticKITTI [60] provides large-scale annotated point clouds based on the KITTI Odometry Benchmark. We retained two sequences (sequences 0 and 7) for the SE-Calib evaluation and others for training the 3-D-SS model Cylinder3D [8]. Both DFF and Cylinder3D were trained with 19 classes, 14 common classes were utilized in calibration, and others were ignored.

3) Hardware and Software: All the evaluations were completed on a laptop equipped with AMD R4800H CPU and NVIDIA GeForce RTX2060 GPU with C++ and Python implementation. Our operation system is Ubuntu20.04 LTS.

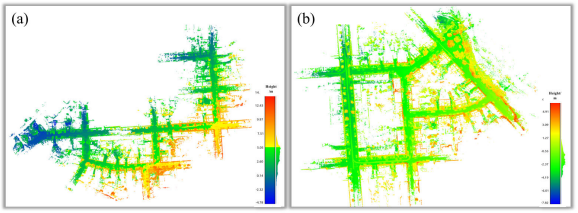

Fig. 5. Point clouds submaps of two study areas. (a) Area I. (b) Area II.

TABLE I
Dataset Description

Parameters	Area I	Area II
Frames	1000	1000
Resolution	$\approx 1240 \times 370$	$\approx 1240 \times 370$
Number of points	119,864,871	120,370,063
Length(m)	≈ 700	≈ 500

4) Hyperparameters: The origin RGB images with approximately size $(h, w) = (370, 1240)$ were cropped to $(h, w) = (320, 1024)$ before fed into 2-D-SED pipeline. Cylinder3D [8] split point clouds into 3-D cylindrical voxel with size $= 480 \times 360 \times 32$ in the cylindrical coordinate system in their approach, where the dimensions indicate the radius r, azimuth θ, and height z. We followed their settings in our experiments. The search step $ssp = 0.7°$ and the minimum search step threshold is one-tenth of the ssp. The spatial and temporal smoothness threshold $(\epsilon_s, \epsilon_t) = (0.35, 0.9)$ in the CCuP. The analysis of hyperparameters settings is detailed in SecV-D.

B. Quantitative Evaluation

To evaluate the feasibility of our method, we take 1000 consecutive scans as studying areas from two different sequences, as shown in Fig. 5. Table I lists the details of the experimental areas. For each scan, we apply the perturbation distributed in $[1°, 2°]$ uniformly to the ground-truth boresight parameters (provided by the calibration file) as initialization, where the sign of perturbation is randomized. Fig. 6 visualizes the solutions of the SE-Calib among test frames, where the first row shows the confidence value and the second row shows the residual between optimized parameters and the ground-truth parameters. As shown in Fig. 6(a), the proposed method reduces the roll, pitch, and yaw residual to 0.165°, 0.147°, and 0.184° in area I; in Fig. 6(b), the mean average residual of roll, pitch, and yaw in area II is 0.265°, 0.213°, and 0.261°, respectively.

Although considerable performance was achieved on most frames, severe miscalibration encounters on some frames. Interpretation of miscalibration is some nasty frames, with confusing occlusion or few unique features as mentioned in Section III-B, infiltrated into calibration frames. To filter out those unreliable frames, two additional tricks are available: 1) removing outliers based on CCuP and 2) selecting the frame with the temporally highest confidence level for every small period of time, called "temporal reliability selection" (TRC). The goals of CCuP and TRC are different. The

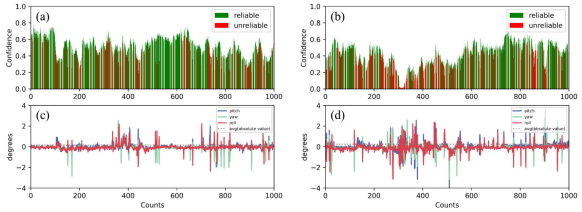

Fig. 6. Calibration results of SE-Calib. (a) and (b) Confidence of calibration results in areas I and II, reliable frames denoted with green histogram and unreliable frames (detected by CCuP) denoted with red histogram. (c) and (d) Boresight parameters error after calibration in areas I and II.

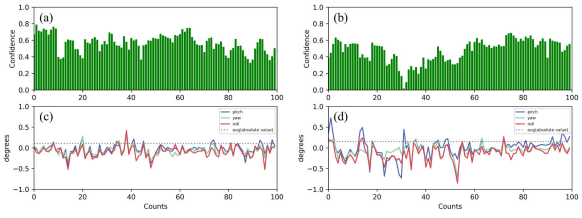

Fig. 7. Calibration results of SE-Calib+TRC. (a) and (b) Confidence of calibration results in areas I and II. (c) and (d) Boresight parameters error after calibration in areas I and II.

CCuP is proposed to eliminate all the unreliable frames from noisy scenes as much as possible (without considering the distribution of reliable frames). While the TRC is to provide time-continuous and reliable calibration results. In urban MMS or AV applications, the extrinsic transform parameters should keep steady for a short period of time (because of the flatness of the road). Therefore, if we could select one "temporal reliable" frame for a short period of time (1 s in our experiment) with TRC or select all reliable frames during a large period of time (100 s in our experiment) with CCuP and then propagate the optimized parameters of those reliable frames to the nearby unreliable frames, a more reliable calibration result can be achieved in the aspect of engineering.

Thus, we test CCuP and TRC to extract reliable results. Fig. 7 visualizes the solutions of the SE-Calib+TRC, where the mean average residual of roll, pitch, and yaw reduced to 0.095°, 0.100°, and 0.129° and 0.166°, 0.121°, and 0.172°, respectively. Comparing Fig. 7 with Fig. 6, all the misregistrations were filtered out and a more accurate solution, with only 0.130° residual across all the dimensions and frames, was achieved, which indicates both robustness and accuracy of our proposed TRC strategy. Also, Table II counts the calibration results of vanilla SE-Calib, SE-Calib + CCuP, and SE-Calib + TRC. It shows that the SE-Calib + CCuP and SE-Calib + TRC achieve similar calibration results with 0.129° and 0.130° average error of the boresight parameters.

To qualitatively evaluate the performance of our method in a more direct way, point-pixel corresponding pairs are checked manually to calculate the reprojection error. Checkpoints should be distinguishable and evenly distributed over the full image, such as trunks, building corners, or vehicles. After manually selecting and projecting checkpoints of point clouds onto the image plane, the offsets between projection pixels and ground-truth pixels are measured, as formatted in

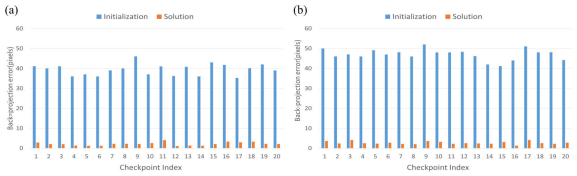

Fig. 8. Error distribution in study sites. (a) Site a. (b) Site b.

the following equation:

$$\text{pixel}_{\text{error}} = \sqrt{(x_{\text{proj}} - x_{gt})^2 + (y_{\text{proj}} - y_{gt})^2} \tag{4}$$

where (x_{gt}, y_{gt}) denotes the ground-truth pixels coordinates and $(x_{\text{proj}}, y_{\text{proj}})$ denotes the projection pixels coordinates of checkpoints on the image plane. The reprojection error of initial transform parameters and the optimized error of two selected sites are counted and shown in Table III. Fig. 8 shows the error distribution in these sites.

C. Comparative Evaluation

We reimplemented two existing calibration methods based on artificial features and tested with identical settings as our baseline, including line feature [22] and edge feature [15]. In the absence of publicly released codes, we developed our best adaptations of these methods and verified them on identical test frames as described in Section IV-A, where empirical results are shown in Table II. The results indicate that the proposed SE-Calib outperforms baselines by a large margin. The CCuP and TRC strategies further enhance our superiority.

D. Calibration Visualization

To visualize the calibration effect, we project the raw point clouds onto the image plane with initial transform parameters and optimized transform parameters, respectively, as shown in Figs. 9 and 10. Following the colorization rules of SemanticKITTI [60], cars are colorized blue and poles are colorized yellow. The alignment between point clouds and images improves noticeably after calibration. More results can be found here.[1][2]

E. Time Performance

Time performance statistics of SE-Calib is reported in Table IV. We utilize the official timing interface *torch.cuda.Event* of PyTorch to track the time cost of 2-D-SED [24] and 3-D-SS [8] inferences. Due to the limited GPU memory (6 GB), the 2-D-SED and 3-D-SS processes are carried out sequentially. Since the lightweight 3-D-SED process is parallel, time consumption is negligible. The total time cost is less than 0.6 s (0.575 s). We can find that the proposed pipeline could perform online even on a mobile device with limited computation and storage capacity.

[1]https://youtu.be/WstVReWA-SA
[2]https://www.bilibili.com/video/BV1sc411H7iS

TABLE II
QUANTITATIVE RESULTS OF SE-CALIB AND BASELINES

Setting	Sequences 0(°)			Sequences 7(°)			Mean(°)
	Roll	Pitch	Yaw	Roll	Pitch	Yaw	
SE-Calib	0.165	0.147	0.184	0.265	0.213	0.261	0.206
SE-Calib+CCuP	0.126	0.103	0.141	0.148	**0.103**	**0.151**	**0.129**
SE-Calib+TRC	**0.095**	**0.100**	**0.129**	**0.166**	0.121	0.172	0.130
Edge feature[15]	1.043	1.029	1.118	1.043	0.952	1.016	1.034
Line feature[22]	0.411	0.374	0.352	0.623	0.611	0.551	0.487

TABLE III
REPROJECTION ERROR WITH INITIAL BORESIGHT PARAMETERS AND CALIBRATED BORESIGHT PARAMETERS

Setting	Initial Error(pixel)			Calibrated Error(pixel)		
	Average	Max	RMSE	Average	Max	RMSE
Site a	39.43	46.05	2.78	2.20	4.02	0.79
Site b	47.02	52.01	2.66	2.74	4.16	0.70

TABLE IV
TIME PERFORMANCE OF THE PROPOSED SE-CALIB ON A LAPTOP EQUIPPED WITH AMD R4800H CPU AND NVIDIA RTX 2060 GPU

Time Performance (ms)	CPU	GPU
2D-SED	N/A	182.6094
3D-SS	N/A	222.8235
lightweight 3D-SED	N/A	N/A
Boresight optimization	169.4541	N/A
Total	574.8870	

Fig. 9. Qualitative results of the SE-Calib in study area I. (a) Projection with the initial boresight parameters. (b) Projection with the optimized parameters.

V. DISCUSSION

A. Comparation of Semantic Edge and Geometric Edge

In the adaptive optimization process, we exploit semantic edge features for calibration. However, existing studies have shown that geometric edge features based on simple distance discontinuities can also perform well in calibration task [15], [18], [33]. To evaluate the importance of semantic edge features in our work, we replace semantic edge features on point clouds with geometric edge features detected by a lightweight method [15]. Due to the lack of semantic information in point clouds, the objective function is modified as projecting edge points onto the image plane and calculating the probability that the projection pixels are classified as an edge in the 2-D-SED network, formatted in the following equation:

$$s(\boldsymbol{R}) = \sum_{f_k \in F_{se}}^{f_k} \boldsymbol{I}_{ij}^{k}[\boldsymbol{K}(\boldsymbol{R} f_k + \boldsymbol{t})]. \tag{5}$$

Fig. 10. Qualitative results of the SE-Calib in study area II. (a) Projection with the initial boresight parameters. (b) Projection with the optimized parameters.

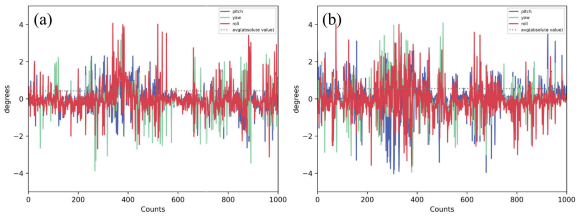

Fig. 11. Calibration results with geometry edge feature. (a) Boresight parameters error in study area I. (b) Boresight parameters error in study area II.

We present the converged solutions based on the same test ground and settings in Fig. 11. A comparison of the results in Fig. 11 with those in Fig. 6 shows that the use of geometric edge features results in inaccurate calibration results. This indicates that semantic information significantly improves the robustness of the edge features.

B. Comparison of Soft Targets and Hard Targets

In the feature extraction step, we store the soft targets for both image and point clouds instead of hard targets. This approach involves tagging each pixel or point with a float vectorized feature that represents the probability (instead of a binary one), where each entry indicates the predicted probability of the corresponding semantic edge feature category. Soft targets are more informative than hard targets as they adaptively weigh the different corresponding pairs, reducing the impact of less determinate semantic edge features and increasing the effect of more determinate ones. To demonstrate our inference that soft targets with advantages in calibration tasks, we replace the soft targets f with hard targets l through a binary mapping function $B(.)$ in Eq. (6)

$$l_i = B(f_i) = \begin{cases} 1, & f_i \geq 0.5 \\ 0, & f_i < 0.5 \end{cases} \tag{6}$$

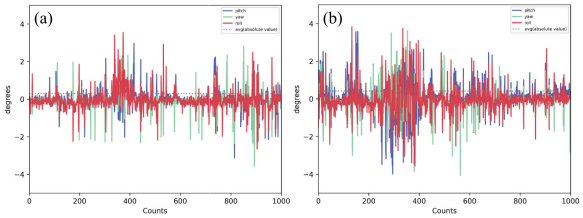

Fig. 12. Calibration results with hard targets. (a) Boresight parameters error in study area I. (b) Boresight parameters error in study area II.

TABLE V
ABLATION STUDIES ON THE SEMANTIC EDGE AND SOFT TARGETS

Setting	Estimation Error(°)			
	Roll	Pitch	Yaw	Mean
SE+ST(proposed)	**0.215**	**0.180**	**0.223**	**0.206**
GE+ST	0.709	0.692	0.580	0.660
SE+HT	0.514	0.495	0.429	0.479

where i means the ith class. Evaluations are conducted and shown in Fig. 12.

A comparison between Figs. 6 and 12 reveals significantly degraded calibration results with hard targets in both sequences, particularly in area II [Fig. 12(b)]. This indicates that hard targets lead to poorer accuracy of the calibration.

For a quantitative assessment of our proposed method versus the alternative methods, we present the average estimation error across two sequences in Table V. "SE" and "GE" are for semantic edge and geometric edge, respectively, and "ST" and "HT" are for soft targets and hard targets, respectively.

C. Analysis of Features Extraction Noise

In most feature-based calibration methods, outlier detection and removal is a necessary process after feature extraction. In [22], detected line features with few adjacent points or shorter than 8 pixels are removed. In [21], detected parking vehicle pairs with small overlapping areas are masked as false positive objects and removed. As the SSRCM rule inherently comes with the error rejection property, we argue that our SE-Calib does not require either the outlier detection or removal block. Theoretical analysis and ablation experiments demonstrate our suppose in this section.

In Eq. (2), for each corresponding point-pixel features pair (f_k, f_{ij}), a high response score exists only when the semantic edge features from different modalities with very similar orientations. Regardless of which side the false detection occurs, either the image part or the point clouds part, the orientations of semantic edge features will diverge, and then, the false corresponding pair will get a low score and contribute little to the global semantic consistency. In other words, with the SSRCM, the semantically inconsistent matching pairs are given lower scores and semantically consistent pairs are given higher scores, which is in substance similar to the robust estimation. To simulate outliers in the feature extraction step, we use random noise distributed in $U[0, 1]$ to mask out patches of the semantic edge feature maps. The contaminated semantic edge feature maps are shown in Fig. 13. Then, we perform

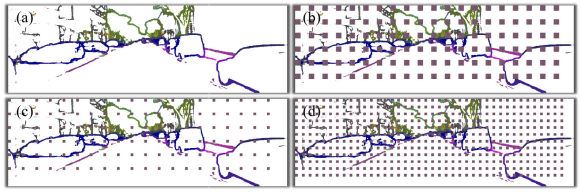

Fig. 13. Examples of contaminated semantic edge map. (a) Clean semantic edge map. (b) Noise patches with a gap of 50 pixels and a side length of 20 pixels. (c) Noise patches with a gap of 50 pixels and a side length of 10 pixels. (d) Noise patches with a gap of 25 pixels and a side length of 10 pixels.

TABLE VI
ABLATION STUDIES ON THE FEATURE EXTRACTION NOISE

Setting		Estimation Error(°)			
		Row	Pitch	Yaw	Mean
$noise = (20, 50)$	SE-Calib	0.453	0.420	0.462	0.445
	+CCuP	0.246	0.191	0.257	0.231
	+TRC	0.217	0.182	0.241	0.213
$noise = (10, 50)$	SE-Calib	0.304	0.271	0.332	0.302
	+CCuP	**0.170**	**0.133**	**0.186**	**0.163**
	+TRC	0.183	0.149	0.208	0.180
$noise = (10, 25)$	SE-Calib	0.433	0.395	0.466	0.431
	+CCuP	0.255	0.192	0.255	0.234
	+TRC	0.230	0.197	0.264	0.230

the optimization process following the identical settings in Section IV-A. The calibration results are shown in Table VI. In Table VI, noise $= (a, b)$ means noise patches with a side length of a pixels and a gap of b pixels.

Comparing the experimental results shown in Tables II and VI, we find that the vanilla SE-Calib achieves accurate results despite the noise and the CCuP and TRC strategies improve the calibration accuracy by a large margin. We speculate that the CCuP and TRC are still powerful to remove unreliable frames even under severe disturbances.

D. Hyperparameters Settings and Sensitive Analysis

The search step ssp in the adaptive optimization process determines the precision of the optimization results, and the smoothness thresholds (ϵ_s, ϵ_t) in the CCuP trick influence the refined results directly. As shown in Fig. 14 and Table VII, setting ssp to a small value (e.g., 0.5) leads to the local optimum that the calibration results of vanilla SE-Calib degrade obviously. On the contrary, a big value (e.g., 0.9) leads to higher error when applying CCuP and TRC tricks. Therefore, we choose the moderate step size $ssp = 0.7$ for both accurate and reliable calibration results.

For the smoothness thresholds (ϵ_t, ϵ_s), we utilize the grid search approach to select eight sets of different values and show the refined calibration results of SE-Calib + CCuP in Table VIII. The "PCT" in Table VIII means the percentage of frames that are reserved as "reliable frames" under our CCuP rules. Our purpose is to balance the percentage of reserved frames and estimation error, so $(\epsilon_s, \epsilon_t) = (0.35, 0.9)$ is adopted.

Another issue about the sensitivity of smoothness thresholds should be considered is: does the CCuP process particularly

TABLE VII
ABLATION STUDIES ON THE SEARCHING STEP SSP

Setting		Estimation Error(°)			
		Row	Pitch	Yaw	Mean
$ssp = 0.5$	SE-Calib	0.235	0.213	0.261	0.236
	+CCuP	0.140	0.106	**0.146**	0.130
	+TRC	**0.128**	0.112	0.148	**0.129**
$ssp = 0.7$	SE-Calib	0.215	0.180	0.223	0.206
	+CCuP	0.137	**0.103**	**0.146**	**0.129**
	+TRC	0.131	0.111	0.151	0.130
$ssp = 0.9$	SE-Calib	0.196	0.154	0.211	0.187
	+CCuP	0.143	0.111	0.156	0.136
	+TRC	0.140	0.112	0.160	0.137

TABLE VIII
ABLATION STUDIES ON THE SPATIAL AND TEMPORAL SMOOTHNESS THRESHOLD (ϵ_s, ϵ_t). ↑ MEANS THAT BIGGER IS BETTER, AND THE ↓ MEANS THAT SMALLER IS BETTER

Setting	Estimation Error(°) ↓				PCT(%) ↑
	Roll	Pitch	Yaw	Mean	
w/o (ϵ_s,ϵ_t)	0.215	0.146	0.194	0.206	**100.0**
$(\epsilon_s,\epsilon_t) = (0.15, 0.9)$	0.146	0.111	0.155	0.137	64.6
$(\epsilon_s,\epsilon_t) = (0.25, 0.9)$	0.143	0.110	0.154	0.136	63.4
$(\epsilon_s,\epsilon_t) = (0.45, 0.9)$	0.129	0.104	0.147	0.129	47.2
$(\epsilon_s,\epsilon_t) = (0.55, 0.9)$	**0.110**	**0.090**	**0.132**	**0.111**	25.7
$(\epsilon_s,\epsilon_t) = (0.35, 0.7)$	0.152	0.113	0.158	0.141	78.6
$(\epsilon_s,\epsilon_t) = (0.35, 0.8)$	0.143	0.106	0.151	0.133	76.0
$(\epsilon_s,\epsilon_t) = (0.35, 0.95)$	0.130	0.099	0.142	0.124	34.0
$(\epsilon_s,\epsilon_t) = (0.35, 0.0)$	0.156	0.115	0.162	0.144	79.6
$(\epsilon_s,\epsilon_t) = (0.0, 0.9)$	0.144	0.112	0.155	0.137	64.8
$(\epsilon_s,\epsilon_t) = (0.35, 0.9)$	0.137	0.103	0.146	0.129	58.0

rely on the temporal smoothness threshold ϵ_s or the spatial smoothness threshold ϵ_t? Especially when sequential consistency is not available for a single pair of laser frame and image, can CCuP still improve the calibration accuracy using spatial smoothness? To verify the issue, we remove the spatial smoothness ϵ_s and the temporal smoothness ϵ_t (set to 0 empirically) and report the refined calibration results in Table VIII. According to the results in Table VIII, only the ϵ_s or ϵ_t threshold could refine the calibration results to an average error of 0.144° and 0.137°, which indicates that the CCuP could still remove unreliable calibration results powerfully without temporal or spatial consistency cue.

The other parameters used in our experiments are relatively easily and reasonably set. The resolution of the RGB images is not uniform, so we cropped the images to $(h, w) = (320, 1024)$ for semantic edge prediction. Cylinder3D [8] divided point clouds into small grids in the cylindrical coordinate system. They claimed that the cylindrical voxel partition met the varying sparsity of driving-scene LiDAR point cloud and balanced point distribution. We followed their settings in our experiments. Since the SE-Calib method could achieve a 0.129° average error at the best case, setting the minimum step threshold to one-tenth of the search step ssp (0.07° in our experiment) is a good balance of accuracy and efficiency.

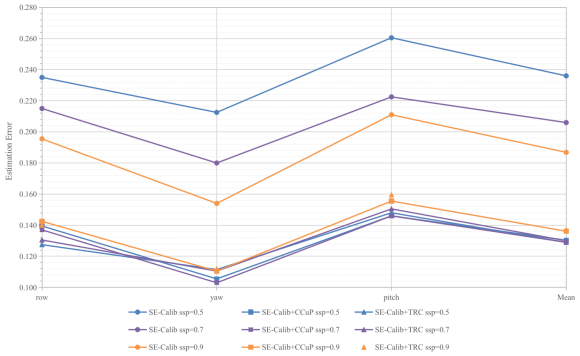

Fig. 14. Calibration results with different search steps ssp.

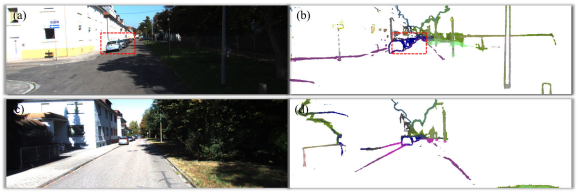

Fig. 15. Examples of failure calibration frames. The first column shows the RGB images and the second column shows the 2-D-SED prediction results. (a) and (b) False feature extraction. (c) and (d) Short of discriminative semantic edge features.

E. Analysis of Failed Calibration Cases

The accuracy of the SE-Calib solution relies heavily on the quality and quantity of semantic edge features extracted from point clouds and images. We analyze two failure scenarios in Fig. 15. In Fig. 15(a), due to the large distance and low resolution, several internal pixels of the continuously parked vehicles are incorrectly detected as boundary pixels. In Fig. 15(b), the SE-Calib fails due to insufficient representation of discriminative and robust semantic edge characteristics.

VI. CONCLUSION

Accurate boresight parameters between the LiDAR and camera are crucial for multimodalities data fusion tasks. In this article, we propose SE-Calib, a novel online calibration method that utilizes semantic edge features to calibrate the boresight parameters between LiDAR and the camera. The main contribution of our method is constructing accurate and robust matching pairs using the semantic edge features. With the proposed SSRCM rules, the optimization process finds the optimal value iteratively, while the reliability check-up process rejects unreliable results. Experiments on the KITTI dataset have shown that SE-Calib achieves accuracy and robustness in urban scenes. We expect that the proposed method could benefit the society of MMS and robotics.

While the results are encouraging, some challenges still remain. As the distribution and resolution of images in other evaluation datasets may be different from the training set, the performance of the 2-D-SED module degrades. Several cut-edge approaches [65], [66] have proven that unsupervised domain adaption (UDA) methods could mitigate the problem caused by domain gap. Also, utilizing a more powerful and lightweight 2-D-SED module is another way to improve the

performance of the 2-D-SED task. Besides, the alternative 3-D-SED process is not end-to-end, which may reduce the efficiency of the calibration task. In future work, we intend to design novel, accurate, and lightweight 2-D-SED/3-D-SED modules to improve the semantic edge feature extraction performance. UDA techniques will be utilized to enhance the SED accuracy in agnostic environments.

REFERENCES

[1] T. Zhou, S. M. Hasheminasab, and A. Habib, "Tightly-coupled camera/LiDAR integration for point cloud generation from GNSS/INS-assisted UAV mapping systems," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 336–356, Oct. 2021.

[2] J. Li et al., "3D mobile mapping with a low cost UAV system," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W8, pp. 127–132, Nov. 2017.

[3] Y. Cong et al., "3D-CSTM: A 3D continuous spatio-temporal mapping method," *ISPRS J. Photogramm. Remote Sens.*, vol. 186, pp. 232–245, Apr. 2022.

[4] T. Shan and B. Englot, "LeGO-LOAM: Lightweight and ground-optimized LiDAR odometry and mapping on variable terrain," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 4758–4765.

[5] J. Zhang and S. Singh, "LOAM: LiDAR odometry and mapping in real-time," in *Proc. Robot., Sci. Syst.*, vol. 2, 2014, pp. 1–9.

[6] X. Yan et al., "2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel: Springer, Oct. 2022, pp. 677–695.

[7] C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Marlet, "Image-to-LiDAR self-supervised distillation for autonomous driving data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9891–9901.

[8] X. Zhu et al., "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9939–9948.

[9] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12697–12705.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[11] Y. Zhu, C. Li, and Y. Zhang, "Online camera-LiDAR calibration with sensor semantic information," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4970–4976.

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

[13] J. Kümmerle and T. Kühner, "Unified intrinsic and extrinsic camera and LiDAR calibration under uncertainties," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6028–6034.

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

[15] J. Levinson and S. Thrun, "Automatic online calibration of cameras and lasers," in *Proc. Robot., Sci. Syst. IX*, Jun. 2013, pp. 1–8.

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

[17] S. Wang et al., "Temporal and spatial online integrated calibration for camera and LiDAR," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 3016–3022.

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

[19] T. Li, J. Fang, Y. Zhong, D. Wang, and J. Xue, "Online high-accurate calibration of RGB+ 3D-LiDAR for autonomous driving," in *Proc. Int. Conf. Image Graph.* Shanghai, China: Springer, Sep. 2017, pp. 254–263.

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

[21] J. Li, B. Yang, C. Chen, R. Huang, Z. Dong, and W. Xiao, "Automatic registration of panoramic image sequence and mobile laser scanning data using semantic features," *ISPRS J. Photogramm. Remote Sens.*, vol. 136, pp. 41–57, Feb. 2018.

[22] X. Zhang, S. Zhu, S. Guo, J. Li, and H. Liu, "Line-based automatic extrinsic calibration of LiDAR and camera," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 9347–9353.

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

[24] Y. Hu, Y. Chen, X. Li, and J. Feng, "Dynamic feature fusion for semantic edge detection," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 782–788.

[25] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2004, pp. 2301–2306.

[26] R. Unnikrishnan and M. Hebert, "Fast extrinsic calibration of a laser rangefinder to a camera," Robot. Inst., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-05-09, 2005.

[27] D. Scaramuzza, A. Harati, and R. Siegwart, "Extrinsic self calibration of a camera and a 3D laser range finder from natural scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2007, pp. 4164–4169.

[28] T. Tóth, Z. Pusztai, and L. Hajder, "Automatic LiDAR-camera calibration of extrinsic parameters using a spherical target," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8580–8586.

[29] V. Fremont, S. A. Rodriguez F., and P. Bonnifait, "Circular targets for 3D alignment of video and LiDAR sensors," *Adv. Robot.*, vol. 26, no. 18, pp. 2087–2113, Dec. 2012.

[30] X. Gong, Y. Lin, and J. Liu, "3D LiDAR-camera extrinsic calibration using an arbitrary trihedron," *Sensors*, vol. 13, no. 2, pp. 1902–1918, Feb. 2013.

[31] S. Debattisti, L. Mazzei, and M. Panciroli, "Automated extrinsic laser and camera inter-calibration using triangular targets," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2013, pp. 696–701.

[32] Y. Ye, B. Zhu, T. Tang, C. Yang, Q. Xu, and G. Zhang, "A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 331–350, Jun. 2022.

[33] Z. Taylor, J. Nieto, and D. Johnson, "Automatic calibration of multi-modal sensor systems using a gradient orientation measure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 1293–1300.

[34] L. Zhou, Y. Ye, T. Tang, K. Nan, and Y. Qin, "Robust matching for SAR and optical images using multiscale convolutional gradient features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[35] K. Irie, M. Sugiyama, and M. Tomono, "Target-less camera-LiDAR extrinsic calibration using a bagged dependence estimator," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2016, pp. 1340–1347.

[36] 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 *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 2053–2059.

[37] E. G. Parmehr, C. S. Fraser, C. Zhang, and J. Leach, "Automatic registration of optical imagery with 3D LiDAR data using statistical similarity," *ISPRS J. Photogramm. Remote Sens.*, vol. 88, pp. 28–40, Feb. 2014.

[38] P. Jiang, P. Osteen, and S. Saripalli, "SemCal: Semantic LiDAR-camera calibration using neural mutual information estimator," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2021, pp. 1–7.

[39] Z. Taylor and J. Nieto, "Motion-based calibration of multimodal sensor arrays," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4843–4850.

[40] Z. Taylor and J. Nieto, "Motion-based calibration of multimodal sensor extrinsics and timing offset estimation," *IEEE Trans. Robot.*, vol. 32, no. 5, pp. 1215–1229, Oct. 2016.

[41] B. D. Corte, H. Andreasson, T. Stoyanov, and G. Grisetti, "Unified motion-based calibration of mobile multi-sensor platforms with time delay estimation," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 902–909, Apr. 2019.

[42] A. Sen, G. Pan, A. Mitrokhin, and A. Islam, "SceneCalib: Automatic targetless calibration of cameras and LiDARs in autonomous driving," 2023, *arXiv:2304.05530*.

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

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

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

[46] A. Kodaira, Y. Zhou, P. Zang, W. Zhan, and M. Tomizuka, "SST-Calib: Simultaneous spatial–temporal parameter calibration between LiDAR and camera," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 2896–2902.

[47] B. Zhu, C. Yang, J. Dai, J. Fan, Y. Qin, and Y. Ye, "R_2FD_2: Fast and robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5606115.

[48] Q.-H. Pham, M. A. Uy, B.-S. Hua, D. T. Nguyen, G. Roig, and S.-K. Yeung, "LCD: Learned cross-domain descriptors for 2D–3D matching," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11856–11864.

[49] M. Feng, S. Hu, M. H. Ang, and G. H. Lee, "2D3D-matchnet: Learning to match keypoints across 2D image and 3D point cloud," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4790–4796.

[50] B. Wang et al., "P2-Net: Joint description and detection of local features for pixel and point matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16004–16013.

[51] J. Li and G. Hee Lee, "DeepI2P: Image-to-point cloud registration via deep classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15960–15969.

[52] S. Ren, Y. Zeng, J. Hou, and X. Chen, "CorrI2P: Deep image-to-point cloud registration via dense correspondence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1198–1208, Mar. 2023.

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

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

[55] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam, "CASENet: Deep category-aware semantic edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5964–5973.

[56] M. Zhen et al., "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13666–13675.

[57] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[58] Z. Hu, M. Zhen, X. Bai, H. Fu, and C.-L. Tai, "JSENet: Joint semantic segmentation and edge detection network for 3D point clouds," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, Aug. 2020, pp. 222–239.

[59] S. Xia and R. Wang, "A fast edge extraction method for mobile LiDAR point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1288–1292, Aug. 2017.

[60] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9297–9307.

[61] B. O. Abayowa, A. Yilmaz, and R. C. Hardie, "Automatic registration of optical aerial imagery to a LiDAR point cloud for generation of city models," *ISPRS J. Photogramm. Remote Sens.*, vol. 106, pp. 68–81, Aug. 2015.

[62] B. Yang and C. Chen, "Automatic registration of UAV-borne sequent images and LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 101, pp. 262–274, Mar. 2015.

[63] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, "Spatio-temporal self-supervised representation learning for 3D point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6535–6545.

[64] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[65] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[66] L. Yi, B. Gong, and T. Funkhouser, "Complete & label: A domain adaptation approach to semantic segmentation of LiDAR point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15363–15373.

Youqi Liao is currently pursuing the bachelor's degree with the School of Geodesy and Geoinformation (SGG), Wuhan University, Wuhan, China.

He finished this work during an internship at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include the mobile mapping system and lightweight robot perception system.

Jianping Li (Member, IEEE) received the M.S. degree in geographic information system (GIS) and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2015 and 2021, respectively.

He is currently a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include 3-D sensing system integration, unmanned aerial vehicle (UAV)/unmanned ground vehicle (UGV) mapping, robot perception, and point cloud data processing.

Shuhao Kang is currently pursuing the bachelor's degree with the School of Geodesy and Geoinformation (SGG), Wuhan University, Wuhan, China.

He is also an Intern at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include 3-D vision and robotics.

Qiang Li received the B.E. and M.E. degrees in remote sensing science and technology from Wuhan University, Wuhan, China, in 2010 and 2012, respectively.

He is currently a Senior Engineer with Shenyang Geotechnical Investigation and Surveying Research Institute Company Ltd., Shenyang, China. His research interests lie in the fields of real-world 3-D construction, stereo laser scanning systems, and spatial–temporal big data processing.

Guifang Zhu received the M.E. degree in geographic information system from Wuhan University, Wuhan, China, in 2013.

She is currently a Senior Engineer with Shenyang Geotechnical Investigation and Surveying Research Institute Company Ltd., Shenyang, China. Her research interests include multisource point cloud processing and production.

Shenghai Yuan received the B.E. and Ph.D. degrees in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2013 and 2019, respectively.

He is currently a Research Fellow with the Center for Advanced Robotics Technology Innovation (CARTIN), Nanyang Technological University. His research interests include the areas of perception, sensor fusion for robust navigation, machine learning, and various autonomous systems.

Zhen Dong (Member, IEEE) received the B.E. and Ph.D. degrees in remote sensing and photogrammetry from Wuhan University, Wuhan, China, in 2011 and 2018, respectively.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interest lies in the field of 3-D computer vision, particularly 3-D reconstruction, scene understanding, point cloud processing, as well as their applications in intelligent transportation systems, digital twin cities, urban sustainable development, and robotics.

Bisheng Yang received the B.S. degree in engineering survey, the M.S. degree, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1996, 1999, and 2002, respectively.

From 2002 to 2006, he held a post-doctoral position at the University of Zürich, Zürich, Switzerland. Since 2007, he has been a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, where he is currently the Director. His main research interests comprise 3-D geographic information systems (GISs), urban modeling, and digital city.

Dr. Yang was a Guest Editor of the *ISPRS Journal of Photogrammetry and Remote Sensing* and *Computers & Geosciences*.