Correspondence-Free Multiview Point Cloud Registration via Depth-Guided Joint Optimisation

Yiran Zhou[1], Yingyu Wang[1], Shoudong Huang[1] and Liang Zhao[2]

Abstract— **Multiview point cloud registration is a fundamental task for constructing globally consistent 3D models. Existing approaches typically rely on feature extraction and data association across multiple point clouds; however, these processes are challenging to obtain global optimal solution in complex environments. In this paper, we introduce a novel correspondence-free multiview point cloud registration method. Specifically, we represent the global map as a depth map and leverage raw depth information to formulate a non-linear least squares optimisation that jointly estimates poses of point clouds and the global map. Unlike traditional feature-based bundle adjustment methods, which rely on explicit feature extraction and data association, our method bypasses these challenges by associating multi-frame point clouds with a global depth map through their corresponding poses. This data association is implicitly incorporated and dynamically refined during the optimisation process. Extensive evaluations on real-world datasets demonstrate that our method outperforms state-of-the-art approaches in accuracy, particularly in challenging environments where feature extraction and data association are difficult.**

I. INTRODUCTION

Point cloud registration is a fundamental task with extensive applications across various domains, including 3D reconstruction, odometry estimation, multi-sensor point cloud fusion, augmented reality, and virtual reality. 3D point clouds are generated from depth data captured by sensors or reconstructed through stereo image matching. Since data is acquired from different sensors or at different times, point clouds are represented in distinct local coordinates. The goal of point cloud registration is to align these disparate point clouds into a unified coordinate system, facilitating the construction of a consistent and comprehensive 3D model.

The most common approach to point cloud registration is pairwise registration, which estimates the transformation needed to align a source point cloud with a target point cloud. However, in applications such as 3D reconstruction, a single pair of point clouds is often insufficient to capture the complete structure of an object. To address this, multiple point clouds must be aligned within a unified coordinate system, a process known as multiview point cloud registration.

Early solutions to the multiview registration problem relied on sequential pairwise registration [1]. However, such approaches suffer from accumulating relative pose errors as the number of frames increases, ultimately failing to produce a globally consistent point cloud. To mitigate this issue, a common approach is to formulate the problem as pose graph optimisation (PGO) [2], [3] or feature-based bundle adjustment (BA) [4]–[9]. PGO-based methods optimise only the poses, while feature-based BA approaches jointly refine both poses and features in the global map. Typically, the joint optimization scheme can lead to more accurate results.

The feature-based BA method for multiview registration typically consists of two key steps. The first step, data association, involves identifying correspondences between features across multiple point clouds. Then, using these known correspondences, the problem is formulated as a nonlinear least squares (NLLS) optimisation, relating both poses and features. When data association is accurate, the NLLS problem can be effectively solved using reliable nonlinear optimisation solvers [10], [11]. However, it is a major challenge to establish reliable data association in complex environments, such as those with highly repetitive textures.

In this paper, we propose a novel correspondence-free multiview point cloud registration method guided by depth information. We leverage the depth information of each point cloud to establish constraints on the global point cloud and poses of individual point clouds. Specifically, we formulate multiview registration as a joint optimisation problem, treating both poses of all point cloud frames and the global map as variables. The key novelty of our approach lies in representing the global map as a depth map, where each point in point clouds is associated with the global depth map via its corresponding poses. In this formulation, data association between multi-frame point clouds is implicitly incorporated through the projection relationship between the point clouds and the depth map and dynamically refined during the optimisation process. Our experimental results demonstrate that the proposed method outperforms state-of-the-art approaches in real-world scenarios, particularly in challenging environments where establishing accurate data association is difficult. The main contributions of our work are summarised as follows:

1) We formulate multiview point cloud registration as a joint optimisation problem, which simultaneously optimises the poses of multiple point clouds and the global map.
2) We represent the global map as a depth map, leveraging raw depth information to guide the optimisation process. This eliminates the need for explicit data association, enabling robustness in complex environments.
3) We provide an analytical derivation of the Jacobian for

Yiran Zhou, Yingyu Wang and Shoudong Huang are with Robotics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia (e-mail: Yiran.Zhou-1@student.uts.edu.au; Yingyu.Wang-1@student.uts.edu.au; Shoudong.Huang@uts.edu.au)

Liang Zhao is with the School of Informatics, The University of Edinburgh, Edinburgh, UK (e-mail: Liang.Zhao@ed.ac.uk).

the proposed optimisation problem, ensuring efficient and accurate problem-solving.

4) Extensive experiments on real-world datasets demonstrate that our method outperforms state-of-the-art approaches in robustness and accuracy.

II. Related Work

A. Pairwise Registration

Pairwise registration is a foundational approach in point cloud registration, widely adopted due to its simplicity and effectiveness in aligning two scans. The earliest method, Iterative Closest Point (ICP) [12], iteratively minimises point-to-point distance but highly sensitive to initial estimates, noise, and outliers, often leading to local minima. To improve robustness and convergence, variants like point-to-plane ICP [13], Generalised ICP [14], and Anisotropic ICP [15] incorporate local point cloud structures and anisotropic covariance modelling. However, these methods remain sensitive to initial pose estimates and struggle in noisy or sparse environments.

To enhance the robustness of these methods, RANSAC [16] is commonly employed. However, its performance deteriorates as the outlier ratios increase. Recently, TEASER [17], a state-of-the-art sequential method for point cloud registration, has been proposed. It follows a sequential approach by estimating pairwise transformations between consecutive frames and incrementally constructing the global point cloud. By leveraging a truncated least-squares optimisation framework, TEASER enhances robustness and efficiency, making it particularly effective under extreme outlier conditions.

Although pairwise registration methods can efficiently and robustly estimate the relative transformation between two point clouds, a single two-frame alignment is often insufficient to capture the complete structure of an object or environment. This limitation restricts their applicability in tasks such as 3D reconstruction.

B. Multiview Registration

Multiview registration aims to optimise the poses of multiple overlapping point cloud frames simultaneously to achieve global consistency. Unlike pairwise registration methods which only consider two frames at a time, multiview registration addresses the global alignment of all frames, reducing accumulated errors and improving overall accuracy.

PGO methods [18], [19] have emerged as a popular approach for multiview registration due to their computational efficiency and scalability. By representing the environment as a graph of keyframes connected by relative pose constraints, these methods efficiently optimise camera trajectories while maintaining real-time performance. However, PGO only optimises the poses of point cloud frames while neglecting the map. This approximation may contribute to suboptimal detail preservation in the registered 3D global point cloud.

In contrast, BA-based methods can jointly optimise poses and maps, fully utilizing observation information. As a result, they are expected to achieve higher precision in reconstruction and more accurate camera pose estimation than PGO-based methods. Traditional feature-based BA methods [4] rely on extracting feature points and establishing constraints by identifying the same feature points across multiple point clouds. However, despite extensive research on feature points extraction [20], [21], reliably detecting and accurately associating feature points in complex scenes with repetitive textures remains challenging, limiting the effectiveness of traditional feature-based BA.

To overcome these limitations, recent research has explored parameterising point clouds into geometric features (e.g., planes or edges) for BA formulation. For example, BALM [7] leverages geometric primitives (e.g., planes, lines) to enhance optimisation stability and reduce computational complexity, while BALM2 [8] and BAREG [9] further improve efficiency by using point clusters and avoiding individual point enumeration. These methods enable the joint optimisation of poses and geometric features, improving accuracy in structured environments. However, their performance deteriorates in unstructured scenes where distinct geometric features are scarce.

Therefore, existing multiview point cloud registration typically depends on explicit feature extraction and data association, which can be unreliable in feature-sparse, noisy environments and unstructured scenes.

C. Joint Optimisation of Poses and Non-feature Map

Several other BA methods also avoid explicit feature extraction and data association by jointly optimising both pose and non-feature-based maps. BAD-SLAM [22] employs a direct BA approach that minimises both reprojection photometric errors and geometric errors. However, directly optimising all pixel points is computationally expensive. To mitigate this, BAD-SLAM adopts an approximation scheme that first optimises only the pose and then updates the surfel representation, which inevitably reduces optimisation accuracy. Additionally, Occupancy-SLAM [23], [24] jointly optimises robot poses and an occupancy grid map to enhance localisation and mapping accuracy. Similarly, [25] introduces an efficient framework to optimise a global occupancy map alongside the coordinate frames of local submaps. Kimera-PGMO [26] jointly optimises poses and mesh representations. While these methods also consider joint optimisation of poses and non-feature-based maps, they are not specifically designed for multiview point cloud registration and employ different non-feature map representations as compared with this paper.

III. Methodology

Our approach considers the joint optimisation of the camera poses and the depth map, leveraging raw depth information without the need for explicit data association. In this section, we will illustrate how the depth observations can be linked to the camera poses to formulate the NLLS problem. Additionally, we obtained an analytical Jacobian derived from the gradient of the depth map, ensuring efficient and accurate optimisation.

A. Task of Multiview Registration and Depth Constraint

The input to multiview point cloud registration is a sequence of point clouds, denoted as $\boldsymbol{P} = \{\boldsymbol{P}_i \mid i = 1, \ldots, N\}$. The task is to estimate their corresponding poses $\boldsymbol{X}^r = \{\boldsymbol{X}_i^r \in SE(3) \mid i = 1, \ldots, N\}$, where $\boldsymbol{X}_i^r = \left[\boldsymbol{t}_i^\top, \boldsymbol{\theta}_i^\top\right]^\top$, so that a consistent global 3D point cloud can be obtained by projecting the individual point clouds into the global coordinate system using these estimated poses. Here, $\boldsymbol{t}_i = [t_i^x, t_i^y, t_i^z]^\top$ represents x-y-z position, and $\theta_i = [\theta_i^x, \theta_i^y, \theta_i^z]^\top$ represents the Euler angles (roll, pitch, yaw) corresponding to rotation matrix $\boldsymbol{R}_i$.

To model the global environment, we first represent it as a 2D depth map $\boldsymbol{D} = [\cdots, \boldsymbol{D}(\boldsymbol{d}_{m,n}), \cdots]$ in the global coordinate, where $\boldsymbol{D}(\boldsymbol{d}_{m,n})$ represents the depth of grid $\boldsymbol{d}_{m,n} (1 \leq m \leq l_m, 1 \leq n \leq l_n)$. The the depth map resolution, s, represents the real-world distance between adjacent grids. The depth value of each grid $\boldsymbol{D}(\boldsymbol{d}_{m,n})$ in the depth map is computed by averaging the depth values of all the points from the 3D point cloud whose x-y coordinates lie within this grid.

Next, the position of the j-th point in the i-th point cloud $\boldsymbol{P}_i$, denoted as $\boldsymbol{p}_{ij}$, can be projected into the global coordinate using its corresponding pose $\boldsymbol{X}_i^r$, i.e.,

$$\boldsymbol{p}'_{ij} = \boldsymbol{R}_i \boldsymbol{p}_{ij} + \boldsymbol{t}_i = [x'_{ij}, y'_{ij}, z'_{ij}]^\top. \tag{1}$$

Assuming that the poses are accurate and the environment is static, the depth value of the projected point located on depth map $\boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)$ should be very closed to its depth measurement in the global coordinate $[\boldsymbol{p}'_{ij}]_z$, which forms the depth constraint. Here,

$$[\boldsymbol{p}'_{ij}]_{xy} = [x'_{ij}, y'_{ij}]^\top, \quad [\boldsymbol{p}'_{ij}]_z = z'_{ij}. \tag{2}$$

The depth measurement $[\boldsymbol{p}'_{ij}]_z$ can be obtained directly from depth sensors (e.g., LiDAR, structured light, and depth cameras) or estimated via a stereo matching algorithm when using a stereo camera, and then projected into the global coordinate.

B. NLLS Formulation

Based on the depth constraint described in Section III-A, we now formulate the NLLS problem to simultaneously optimise the robot poses and the depth map. The state vector of the proposed problem is

$$\boldsymbol{X} = [(\boldsymbol{X}^r)^\top, \boldsymbol{D}^\top]^\top, \tag{3}$$

where

$$\begin{aligned} \boldsymbol{X}^r &= \left[(\boldsymbol{X}_1^r)^\top, \cdots, (\boldsymbol{X}_N^r)^\top\right]^\top, \\ \boldsymbol{D} &= \left[\boldsymbol{D}\left(\boldsymbol{d}_{1,1}\right), \cdots, \boldsymbol{D}\left(\boldsymbol{d}_{l_m,l_n}\right)\right]^\top. \end{aligned} \tag{4}$$

The objective function is defined as

$$f(\boldsymbol{X}) = w_D f^D(\boldsymbol{X}) + w_S f^S(\boldsymbol{X}), \tag{5}$$

where $f^D(\boldsymbol{X})$ and $f^S(\boldsymbol{X})$ represent the depth constraint term and the smoothing term, respectively. w_D and w_S denotes their corresponding weights.

1) Depth Constraint Term: By the local-to-global projection in (1), all points in the point cloud sequence $\boldsymbol{P}$ can be projected to the depth map $\boldsymbol{D}$ to compute the difference in depth values to formulate the NLLS problem, i.e., minimise

$$f^D(\boldsymbol{X}) = \sum_{i=1}^{N} \sum_{j} \left\| [\boldsymbol{p}'_{ij}]_z - \boldsymbol{D}\left(\frac{[\boldsymbol{p}'_{ij}]_{xy}}{s}\right) \right\|^2. \tag{6}$$

Since $[\boldsymbol{p}'_{ij}]_{xy}/s$ may lie at any position on depth map $\boldsymbol{D}$ rather than on a discretised grid, its depth value $\boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)$ can be approximated by bilinear interpolation using depth values of the four neighbouring grids around it. Suppose $[\boldsymbol{p}'_{ij}]_{xy}/s$ locates within four grids, $\boldsymbol{d}_{m,n}, \boldsymbol{d}_{m+1,n}, \boldsymbol{d}_{m,n+1}, \boldsymbol{d}_{m+1,n+1}$, the depth value of $[\boldsymbol{p}'_{ij}]_{xy}/s$ can be calculated by

$$\boldsymbol{D}\left(\frac{[\boldsymbol{p}'_{ij}]_{xy}}{s}\right) = \boldsymbol{H} \begin{bmatrix} \boldsymbol{D}(\boldsymbol{d}_{m,n}) \\ \boldsymbol{D}(\boldsymbol{d}_{m+1,n}) \\ \boldsymbol{D}(\boldsymbol{d}_{m,n+1}) \\ \boldsymbol{D}(\boldsymbol{d}_{m+1,n+1}) \end{bmatrix} \tag{7}$$

where $\boldsymbol{H}$ denotes the bilinear interpolation coefficients.

By using the depth constraint term, the sequence of point clouds $\boldsymbol{P}$, their corresponding poses $\boldsymbol{X}^r$, and the depth map $\boldsymbol{D}$ are linked together.

2) Smoothing Term: To improve the robustness and convergence of our method, we introduce a smoothing term by penalising large variations between neighbouring grids, ensuring that depth transitions smoothly across the map, i.e.,

$$\begin{aligned} f^S(\boldsymbol{X}) = &\sum_{m=1}^{l_m-1} \sum_{n=1}^{l_n-1} \left\| \begin{bmatrix} \boldsymbol{D}(\boldsymbol{d}_{m,n}) - \boldsymbol{D}(\boldsymbol{d}_{m+1,n}) \\ \boldsymbol{D}(\boldsymbol{d}_{m,n}) - \boldsymbol{D}(\boldsymbol{d}_{m,n+1}) \end{bmatrix} \right\|^2 \\ &+ \sum_{n=1}^{l_n-1} \left\| \boldsymbol{D}(\boldsymbol{d}_{l_m,n}) - \boldsymbol{D}(\boldsymbol{d}_{l_m,n+1}) \right\|^2 \\ &+ \sum_{m=1}^{l_m-1} \left\| \boldsymbol{D}(\boldsymbol{d}_{m,l_n}) - \boldsymbol{D}(\boldsymbol{d}_{m+1,l_n}) \right\|^2. \end{aligned} \tag{8}$$

C. Iterative Solution and Analytical Jacobian

1) Iterative Solution: Our NLLS formulation in (5) seeks $\boldsymbol{X}$ such that

$$f(\boldsymbol{X}) = \|F(\boldsymbol{X})\|_{\boldsymbol{W}}^2 = F(\boldsymbol{X})^\top \boldsymbol{W} F(\boldsymbol{X}) \tag{9}$$

is minimised. This formulation can be solved by the Gauss-Newton method. The update vector $\boldsymbol{\Delta}$ in each iteration is the solution to

$$\boldsymbol{J}^\top \boldsymbol{W} \boldsymbol{J} \boldsymbol{\Delta} = -\boldsymbol{J}^\top \boldsymbol{W} F(\boldsymbol{X}) \tag{10}$$

where $\boldsymbol{J}$ is the Jacobian $\partial F(\boldsymbol{X})/\partial X$.

2) Analytical Jacobian: We now derive the analytical Jacobian of our NLLS formulation to accelerate algorithm convergence and enhance robustness.

The Jacobian matrix $\boldsymbol{J}$ consists of three parts: the Jacobian of the depth constraint term w.r.t. the poses $\boldsymbol{J}_P$, the Jacobian of the depth constraint term w.r.t. the depth map $\boldsymbol{J}_D$, and the Jacobian of the smoothing term w.r.t. the depth map $\boldsymbol{J}_S$.

For the depth constraint term $f^D(\boldsymbol{X})$, the Jacobian w.r.t. the pose $\boldsymbol{X}_i^r$ is give by

$$\begin{aligned} \boldsymbol{J}_P &= \frac{\partial\left([\boldsymbol{p}'_{ij}]_z - \boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)\right)}{\partial \boldsymbol{X}_i^r} \\ &= \frac{\partial[\boldsymbol{p}'_{ij}]_z}{\partial \boldsymbol{X}_i^r} - \frac{\partial \boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)}{\partial([\boldsymbol{p}'_{ij}]_{xy}/s)} \cdot \frac{\partial[(\boldsymbol{p}'_{ij}]_{xy}/s)}{\partial \boldsymbol{X}_i^r}. \end{aligned} \tag{11}$$

We can first calculate $\partial(\boldsymbol{p}'_{ij}/s)/\partial \boldsymbol{X}_i^r$ as

$$\frac{\partial(\boldsymbol{p}'_{ij}/s)}{\partial \boldsymbol{X}_i^r} = \left[\frac{\partial(\boldsymbol{p}'_{ij}/s)}{\partial \boldsymbol{t}_i^r}, \frac{\partial(\boldsymbol{p}'_{ij}/s)}{\partial \boldsymbol{\theta}_i}\right] = \frac{1}{s} \cdot [\boldsymbol{I}_{3*3}, \boldsymbol{R}'_i \boldsymbol{p}_{ij}], \tag{12}$$

where $\boldsymbol{R}'_i$ represents the derivative of $\boldsymbol{R}_i$ w.r.t. the orientation. Then we can obtain

$$\begin{aligned} \frac{\partial([\boldsymbol{p}'_{ij}]_{xy}/s)}{\partial \boldsymbol{X}_i^r} &= \frac{1}{s} \cdot [\boldsymbol{I}_{3*3}, \boldsymbol{R}'_i \boldsymbol{p}_{ij}]_{xy}, \\ \frac{\partial[\boldsymbol{p}'_{ij}]_z}{\partial \boldsymbol{X}_i^r} &= [\boldsymbol{I}_{3*3}, \boldsymbol{R}'_i \boldsymbol{p}_{ij}]_z, \end{aligned} \tag{13}$$

where $[\cdot]_{xy}$ and $[\cdot]_z$ are defined in (2).

In (11), $\partial \boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)/\partial([\boldsymbol{p}'_{ij}]_{xy}/s)$ can be considered as the gradient of the depth map at point $[\boldsymbol{p}'_{ij}]_{xy}/s$, which can be approximated by the bilinear interpolation of the gradients of the depth at the four adjacent grid $\nabla \boldsymbol{D}(\boldsymbol{d}_{(m,n)}), \cdots, \nabla \boldsymbol{D}(\boldsymbol{d}_{(m+1,n+1)})$ around $[\boldsymbol{p}'_{ij}]_{xy}$ as

$$\frac{\partial \boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)}{\partial([\boldsymbol{p}'_{ij}]_{xy}/s)} = \boldsymbol{H} \begin{bmatrix} \nabla \boldsymbol{D}(\boldsymbol{d}_{m,n}) \\ \nabla \boldsymbol{D}(\boldsymbol{d}_{m+1,n}) \\ \nabla \boldsymbol{D}(\boldsymbol{d}_{m,n+1}) \\ \nabla \boldsymbol{D}(\boldsymbol{d}_{m+1,n+1}) \end{bmatrix} \tag{14}$$

where the gradient of depth map $\boldsymbol{D}$ at all the grid $\nabla \boldsymbol{D}$ can be easily calculated from the depth map.

The Jacobian of depth constraint term w.r.t. all the grids of depth map $\boldsymbol{J}_D$ can be easily calculated as

$$\begin{aligned} \boldsymbol{J}_D &= -\frac{\partial \boldsymbol{D}([\boldsymbol{p}'_{ij}]_{xy}/s)}{\partial\left[\cdots, \boldsymbol{D}(\boldsymbol{d}_{m,n}), \cdots, \boldsymbol{D}(\boldsymbol{d}_{m+1,n+1}), \cdots)\right]^\top} \\ &= -[0, \cdots, \boldsymbol{H}, \cdots, 0]. \end{aligned} \tag{15}$$

Finally, it is easy to find that the Jacobian of the smoothing term w.r.t. the depth map $\boldsymbol{J}_S$ is equal to a constant matrix that consists of 1, -1, and 0.

IV. EXPERIMENTS

A. Baseline

To evaluate our method's effectiveness, we compare it against five state-of-the-art approaches:

1) Pairwise registration methods: TEASER [17] is one of the most robust and efficient pairwise registration methods. To further enhance the accuracy of TEASER, we incorporate ICP into TEASER for fine registration, denoted as T+ICP. To extend pairwise registration to the multiview registration problem, we apply sequential registration across multiple frames to estimate their poses and refine the final global point cloud.

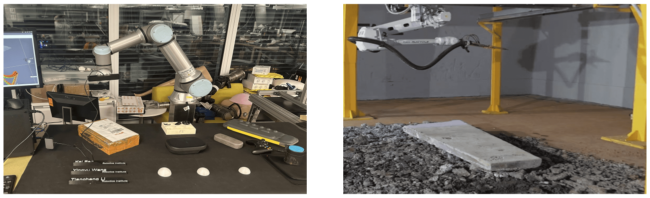
(a) Laboratory Environment (b) Industrial Environment

Fig. 1: The environmental setup for data collection includes both laboratory and industrial scenes. The laboratory environment measures 2.4∗1.2∗0.7 (m), and industrial environment measures 7.0∗6.0∗3.0 (m).

2) Integrating TEASER with batch optimisation: Sequential multiple executions of TEASER in the multiview point cloud registration task inevitably lead to pose error accumulation, resulting in an inconsistent global point cloud map. To address this issue, additional batch optimisation is required to enhance global accuracy across multiple frames. We integrate TEASER with two batch optimisation strategies:

- T+PGO: We apply pose graph optimisation to refine the global alignment, treating the results from TEASER as relative pose measurements.
- T+BA: We extract feature points from point clouds to formulate a feature-based bundle adjustment problem, using the results from sequential TEASER execution as the initial guess.

3) Bundle adjustment methods based on alternative feature representations: BALM2 [8] is the state-of-the-art method in this category, leveraging planar features rather than feature points, as used in T+BA, to construct bundle adjustment.

B. Experimental Datasets and Setup

Our experiments involve two distinct self-collected datasets: a laboratory dataset, captured in a controlled laboratory environment with three scenes (Scene 1-3) and a industrial dataset, recorded in a less structured environment, including three scenes (Scene 4-6) as well. Fig. 1 illustrates the experimental setups for both environments.

1) Laboratory dataset: We collected the dataset with reliable ground truth using a ZED 2 camera mounted on a UR16 robotic arm. The depth accuracy of the camera is less than 1% of the measured depth, and the pose translation errors of UR16 are within 0.05 mm. To obtain accurate ground truth poses, we captured multiple images of a checkerboard from different positions and orientations while simultaneously recording the end-effector poses of the robotic arm at each position. The hand-eye calibration algorithm [27] was then applied to compute the transformation between the camera and the end-effector frame, ensuring accurate alignment with the robotic arm's frame. In addition, the ground truth of the global point cloud is obtained by transforming local point clouds into a global coordinate system using known ground truth poses of depth camera.

2) Industrial dataset: The industrial experiments involved more complex experiments. The setup featured a truss system securely mounted with a structured-light camera, SEIZET

TABLE I: Poses Accuracy of Different Methods Evaluated by Laboratory Dataset

Scene	Metric	TEASER	T+ICP	T+BA	T+PGO	BALM2	Ours
1	MAE((Trans/m))	0.0141	0.0113	0.0083	0.0135	0.0333	0.0068
	RMSE((Trans/m))	0.0202	0.0152	0.0139	0.0206	0.0527	0.0099
	MAE(Rot/rad)	0.0239	0.0118	0.0159	0.0215	0.0446	0.0146
	RMSE(Rot/rad)	0.0349	0.0159	0.0238	0.0345	0.0944	0.0217
2	MAE((Trans/m))	0.0114	0.0124	0.0169	0.0115	0.0413	0.0075
	RMSE((Trans/m))	0.0144	0.0165	0.0292	0.0133	0.0655	0.0092
	MAE(Rot/rad)	0.0111	0.0119	0.0295	0.0229	0.1167	0.0116
	RMSE(Rot/rad)	0.0145	0.0153	0.0479	0.0265	0.1759	0.0151
3	MAE((Trans/m))	0.0392	0.0482	0.0579	0.0643	0.0377	0.0139
	RMSE((Trans/m))	0.6266	0.0655	0.1260	0.0915	0.0638	0.0194
	MAE(Rot/rad)	0.0302	0.2421	0.2093	0.2137	0.0327	0.0141
	RMSE(Rot/rad)	0.0384	0.6266	0.5797	0.3451	0.0524	0.0211

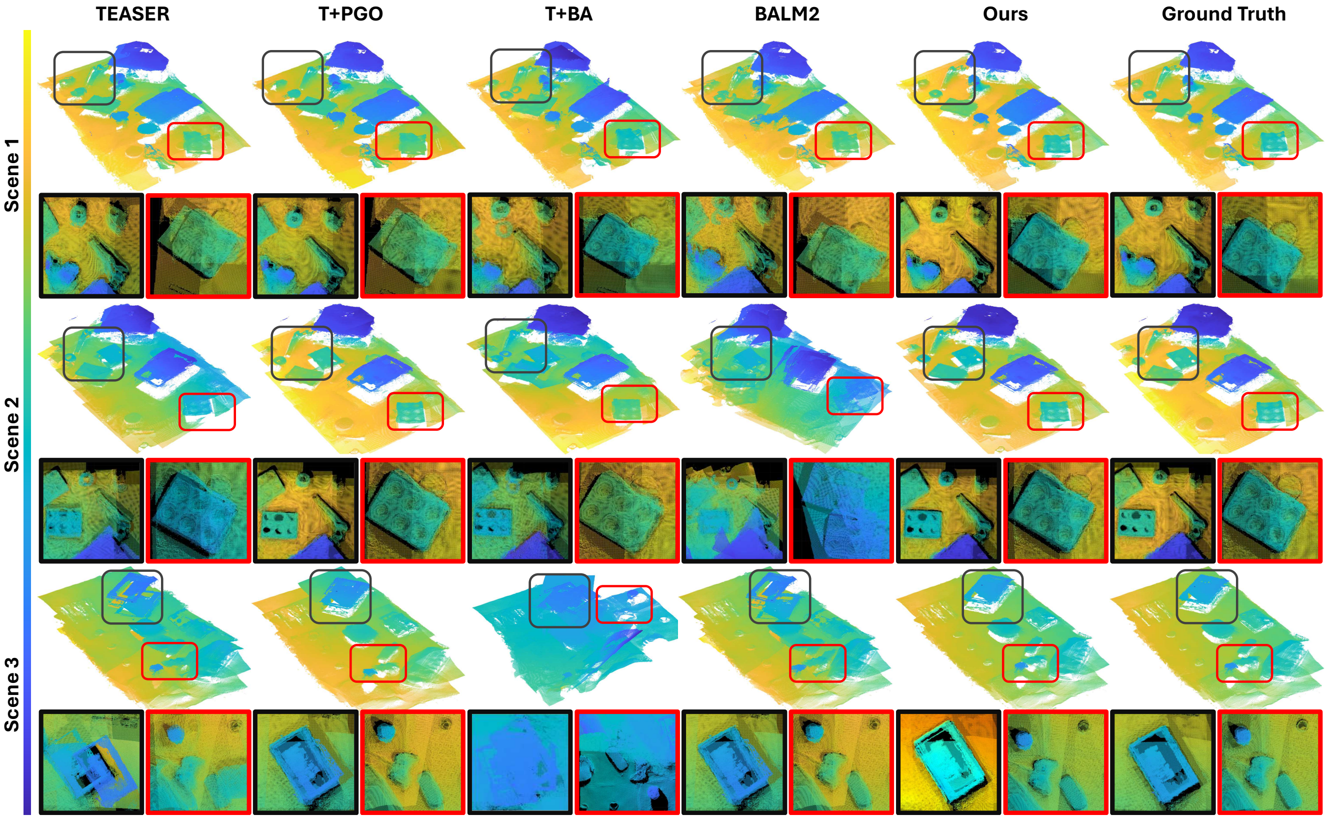

Fig. 2: Global point clouds registered by different methods using the laboratory dataset. The areas highlighted by red and black boxes show detailed figures to enhance the comparison between our and other methods.

SP1000, which provides a depth accuracy of 0.32 mm at a working distance of 3000 mm. The system moves in the X-Y direction only, which is driven by two motors, ensuring a constant height throughout data collection.

The objects captured in this dataset included crushed stone, steel slag, and various types of scrap. The dataset consists of industrial scenes characterised by more complex terrain and less controlled conditions to evaluate performance across different geometric structures. Including a steel slab with a planar surface, a steel coil with a curved surface, and a medium-thick plate with multiple planar surfaces, while keeping the ground features unchanged.

Unlike the laboratory dataset, this dataset does not include ground truth. To facilitate the evaluation of algorithm accuracy, three markers on the floor were strategically placed and fixed within Scenes 4-6. Specifically, one marker was placed at the origin of the global frame, while the other two were positioned at the farthest locations along the X-Y direction. The known distances between these markers serve as a reference for accuracy evaluation.

C. Evaluation of Pose Accuracy

We first quantitatively evaluate the accuracy of the estimated poses using the laboratory dataset, where ground truth of poses are available. Table I presents the quantitative accuracy results across different state-of-the-art methods. We

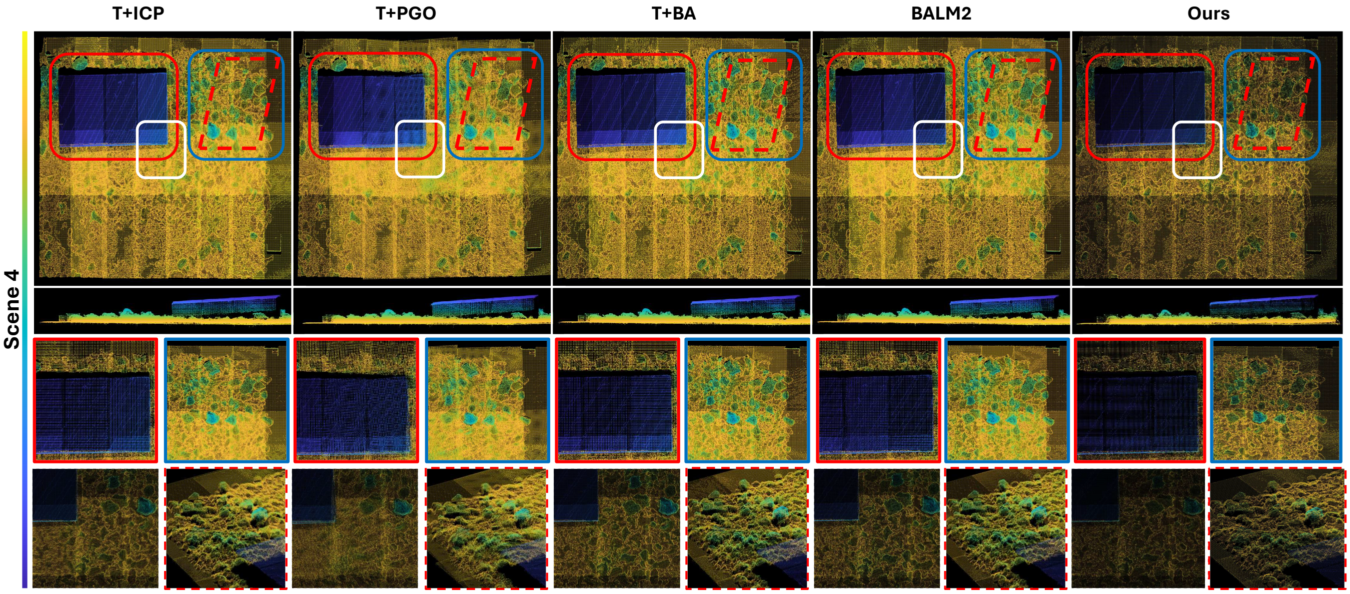

Fig. 3: Comparison of performance among different methods applied to industrial dataset Scene 4.

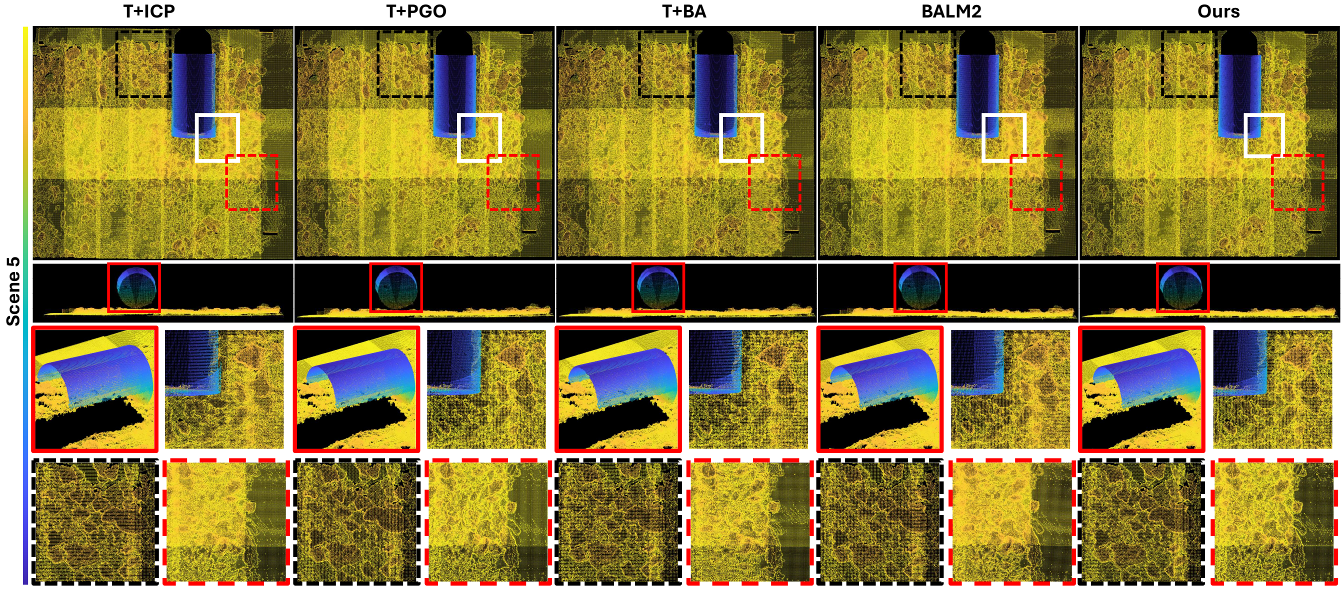

Fig. 4: Comparison of performance among different methods applied to industrial dataset Scene 5.

evaluate pose accuracy using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for both translation (in metres) and orientation (in radians). The best-performing values are highlighted in red, and the second-best in blue.

As shown in Table I, our method achieves the lowest errors across the majority of metrics. In Scene 1, while T+ICP attains the best orientation accuracy, it performs significantly worse in translation against ours. A similar trend is observed in Scene 2, where TEASER achieves better orientation accuracy but exhibits higher translation errors than ours. This demonstrates that some methods excel in one metric but fail to maintain overall accuracy.

A noteworthy observation is that T+BA does not perform well on the laboratory dataset and even underperforms compared to TEASER on several metrics. This is partly because the dataset covers a small scene size with a limited number of frames, resulting in a relatively low accumulation of errors from the sequential execution of TEASER. Consequently, the impact of global optimisation in T+BA is less significant. More importantly, the laboratory dataset contains many regions with repetitive textures, making it difficult for T+BA to establish accurate data association between frames. This challenge directly affects the effectiveness of feature-based bundle adjustment, ultimately reducing its registration accuracy. These results highlight the limitations of T+BA in environments with repetitive structures, where unreliable feature correspondences negatively impact its performance. In contrast, our method does not require feature extraction or explicit data association, allowing it to avoid these limitations. As a result, it remains robust in environments with

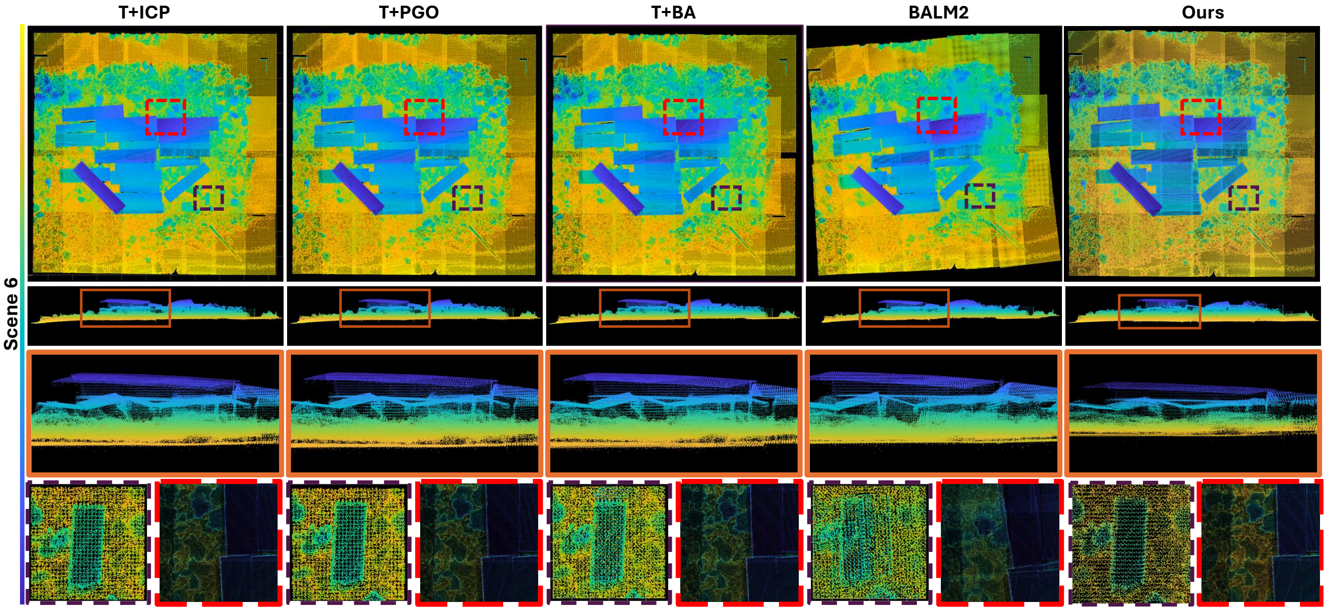

Fig. 5: Comparison of performance among different methods applied to industrial dataset Scene 6.

repetitive textures, maintaining high registration accuracy regardless of scene structure. Additionally, BALM2 does not achieve good results on the laboratory dataset for most metrics. This is because BALM2 relies on planar feature extraction and constructs planar constraints to formulate the bundle adjustment problem. However, in this scenario, many objects are not entirely planar, limiting its effectiveness. In contrast, our approach eliminates the need for feature extraction by associating the point cloud with the global depth map, effectively avoiding this issue.

D. Evaluation of Global Point Cloud Quality

We first qualitatively assess the map quality by visualising the registration results of the 3D global point cloud using both the laboratory dataset and the industrial dataset.

The 3D global point clouds registered by different methods using the laboratory dataset are visualised in Fig. 2. The areas highlighted by red and black rectangles illustrate that our method achieves superior results compared to the others. It can be observed that our method produces a reconstruction closest to the ground truth, with clearer and sharper contour details and well-defined holes. In contrast, the global point cloud registrations produced by TEASER and T+PGO exhibit fuzzy edges and alignment errors across regions. This demonstrates two key issues: first, the sequential execution of TEASER in multiview registration leads to error accumulation, impacting overall accuracy. Second, PGO focuses solely on pose refinement while ignoring the global map, resulting in registered point cloud maps with poorer detail preservation. These findings highlight the importance of jointly optimising both poses and the global map in multiview registration to achieve a more accurate and consistent reconstruction. While BALM2 generally performs well, it struggles in Scene 2 due to the scarcity of high-quality planar features. Notably, T+BA outperforms TEASER in Scene 1 and Scene 2 but fails in Scene 3, likely due to incorrect data association. This highlights the limitations of traditional feature-based BA approaches in complex scenes where feature correspondences become unreliable.

TABLE II: Marker Distances Errors on Industrial dataset

Scene	Direction	TEASER	T+PGO	T+BA	BALM2	Ours
4	Y (m)	0.0479	0.0198	0.0327	0.0339	0.0300
	X (m)	0.0343	0.0059	0.0607	0.0534	0.0152
5	Y (m)	0.0369	0.0246	0.0414	0.0317	0.0226
	X (m)	0.0429	0.0413	0.0530	0.0371	0.0289
6	Y (m)	0.0768	0.0105	0.0139	0.0451	0.0091
	X (m)	0.0186	0.0096	0.0822	0.0536	0.0367

We also qualitatively evaluate map quality using the industrial datasets characterised by less structured environments. As shown in Fig. 3, Fig. 4 and Fig. 5, our method yields a darker global map, indicating higher registration accuracy in overlaps and a more compact reconstruction. Detailed views show sharp contours, preserving fine structures and complex details like scrap geometries and irregular slag edges. For example, in Fig. 3, the red dashed box highlights a steel slag region with distinct colour gradients and sharp edges, while the rectangular billet's upper edge is precisely aligned. In contrast, T+ICP and T+PGO suffer from noticeable blurring and splicing mismatches, with T+PGO showing significant edge misalignment. Similarly, in Fig. 4, the white solid box highlights the edge of the steel coil, where T+ICP shows similar misalignment. Although T+BA and BALM2 improve detail preservation, localised distortions and blurring remain, as highlighted by the blue boxes and red dashed in Fig. 3 and Fig. 4. In contrast, the global point cloud at the top and the deep purple dashed box in Fig. 5 show that T+BA has one misaligned frame, while BALM2 suffers even more misalignment.

Finally, we leverage prior information on marker distances in the industrial dataset to compute surface distance errors, providing a quantitative measure of global point cloud registration quality. As shown in Table II, our method achieves either the best or second-best performance across all metrics. While T+PGO and TEASER slightly outperform our method in certain metrics, the qualitative results in Fig. 3, Fig. 4 and Fig. 5 reveal inconsistencies in the overlaps. This suggests that these frames may suffer from large orientation errors or frames without markers may have inaccurate poses. Therefore, our method demonstrates the most reliable global point cloud registration performance on the industrial dataset.

E. Time Consumption

We further compare the registration runtime with BALM2, T+BA, and T+PGO using Scene 6, which consists of 21 frames of point clouds. For our method, under a resolution of 0.05 m, the number of optimisation iterations until convergence is 15, and the runtime per iteration is approximately 3–5 seconds. In comparison, BALM2 completes the optimisation in around 5 seconds, while T+BA requires 45 seconds in total. Although PGO's pose-only optimization is inherently fast, obtaining multi-frame relative poses using TEASER is significantly more time-consuming, averaging 10 seconds per relative pose computation.

V. Conclusion

This paper formulates the multiview point cloud registration task as a correspondence-free bundle adjustment problem. The key novelty lies in implicitly associating point clouds with a depth map, eliminating the need for explicit feature extraction and data association. As a result, our method avoids the errors commonly encountered in feature-based BA approaches due to insufficient feature extraction or incorrect data association in low-texture, repetitive-texture, or highly unstructured scenarios. We evaluate our method on two self-collected datasets, covering both laboratory and industrial environments, and demonstrate that it outperforms state-of-the-art algorithms, particularly in challenging scenarios where feature extraction and association across multiple frames are difficult.

References

[1] O. D. Faugeras and M. Hebert, "The representation, recognition, and locating of 3-d objects," *The international journal of robotics research*, vol. 5, no. 3, pp. 27–52, 1986.

[2] E. Mendes, P. Koch, and S. Lacroix, "Icp-based pose-graph slam," in *2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2016, pp. 195–200.

[3] H. Wang, Y. Liu, Z. Dong, Y. Guo, Y.-S. Liu, W. Wang, and B. Yang, "Robust multiview point cloud registration with reliable pose graph initialization and history reweighting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9506–9515.

[4] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*. Springer, 2000, pp. 298–372.

[5] F. Dellaert and M. Kaess, "Square root sam: Simultaneous localization and mapping via square root information smoothing," *The International Journal of Robotics Research*, vol. 25, no. 12, pp. 1181–1203, 2006.

[6] L. Zhao, S. Huang, Y. Sun, L. Yan, and G. Dissanayake, "Parallaxba: bundle adjustment using parallax angle feature parametrization," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 493–516, 2015.

[7] Z. Liu and F. Zhang, "Balm: Bundle adjustment for lidar mapping," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3184–3191, 2021.

[8] Z. Liu, X. Liu, and F. Zhang, "Efficient and consistent bundle adjustment on lidar point clouds," *IEEE Transactions on Robotics*, 2023.

[9] H. Huang, Y. Sun, J. Wu, J. Jiao, X. Hu, L. Zheng, L. Wang, and M. Liu, "On bundle adjustment for multiview point cloud registration," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8269–8276, 2021.

[10] S. Agarwal, K. Mierle *et al.*, "Ceres solver: Tutorial & reference," *Google Inc*, vol. 2, no. 72, p. 8, 2012.

[11] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," *Georgia Institute of Technology, Tech. Rep*, vol. 2, p. 4, 2012.

[12] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[13] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.

[14] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp." in *Robotics: science and systems*, vol. 2, no. 4. Seattle, WA, 2009, p. 435.

[15] A. L. Pavlov, G. W. Ovchinnikov, D. Y. Derbyshev, D. Tsetserukou, and I. V. Oseledets, "Aa-icp: Iterative closest point with anderson acceleration," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3407–3412.

[16] A. P. Bustos and T.-J. Chin, "Guaranteed outlier removal for point cloud registration with correspondences," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2868–2882, 2017.

[17] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[18] F. Lu and E. Milios, "Globally consistent range scan alignment for environment mapping," *Autonomous robots*, vol. 4, pp. 333–349, 1997.

[19] K. Pulli, "Multiview registration for large data sets," in *Second international conference on 3-d digital imaging and modeling (cat. no. pr00062)*. IEEE, 1999, pp. 160–168.

[20] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *2008 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2008, pp. 3384–3391.

[21] F. Tombari, S. Salti, and L. Di Stefano, "Performance evaluation of 3d keypoint detectors," *International Journal of Computer Vision*, vol. 102, no. 1, pp. 198–220, 2013.

[22] T. Schops, T. Sattler, and M. Pollefeys, "Bad slam: Bundle adjusted direct rgb-d slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 134–144.

[23] L. Zhao, Y. Wang, and S. Huang, "Occupancy-SLAM: Simultaneously Optimizing Robot Poses and Continuous Occupancy Map," in *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022.

[24] Y. Wang, L. Zhao, and S. Huang, "Occupancy-slam: An efficient and robust algorithm for simultaneously optimizing robot poses and occupancy map," *arXiv preprint arXiv:2502.06292*, 2025.

[25] ——, "Grid-based submap joining: An efficient algorithm for simultaneously optimizing global occupancy map and local submap frames," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 10 121–10 128.

[26] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From slam to spatial perception with 3d dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[27] F. C. Park and B. J. Martin, "Robot sensor calibration: solving ax= xb on the euclidean group," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 5, pp. 717–721, 1994.