The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

A novel CAD-aided coarse-to-fine framework of RGBD-to-point clouds registration

Mengchi Ai[1], Mohamed Elhabiby[1], Naser El-Sheimy[2]

[1]Micro Engineering Tech. Inc., Calgary, AB, Canada, mengchi.ai@microengineering.ca, elhabiby@microengineering.ca
[2]Dept. of Geomatics Engineering, University of Calgary, 2500 University Dr NW, Calgary, AB T2N 1N4, Canada – elsheimy@ucalgary.ca

Keywords: RGBD-to-point clouds registration; Multi-modal; data enhancement; CAD Models; Indoor environments; Coarse-to-fine matching, graph matching.

Abstract
Accurate registration between RGB-D images and point clouds is a critical task for various indoor applications. Estimating the relative pose by aligning the sensor frame with indoor 3D point clouds significantly enhances environmental perception and scene understanding. Existing research primarily focuses on cross-modal feature association through traditional unsupervised methods or supervised learning-based approaches. However, these methods often rely on strong assumptions, such as the availability of an initial pose or substantial overlap between the RGB-D images and the target point clouds. Moreover, the quality of registration is highly sensitive to the density and completeness of the point clouds. To address these limitations, this paper presents a novel coarse-to-fine registration framework with the aid of CAD models. First, a data enhancement process is introduced using the Scan2CAD method to replace functional objects (e.g., chairs and tables) with CAD models, improving semantic and quality consistency. Second, a geometry-aware graph matching is computed to identify regions of interest (ROI) within the point cloud map and estimate the initial pose of the RGBD sensor. Finally, an iterative fine matching using cross-modal is introduced to refine the initial estimated pose. Experimental validation on the ScanNet dataset demonstrates that the proposed framework achieves robust and accurate registration between RGB-D images and 3D point clouds.

1. Introduction

With the rapid advancement of indoor robotics, mobile mapping systems, and related technologies, estimating the relative pose between real-time sensor perception data and pre-built indoor maps has become as a fundamental challenge in both academic research and industry applications (Asl Sabbaghian Hokmabadi et al., 2023). RGB-D sensors, which capture synchronized color and depth information, have emerged as the dominant sensing for indoor environments due to their ability to simultaneously acquire semantic and geometric information. A core task in this domain is determining the relative pose, comprising rotation and translation matrix, between an RGB-D sensor and a pre-built point clouds map of the environment. This paper addresses the problem of estimating the sensor-to-map relative pose, with a particular focus on the challenges associated with cross-modal registration. These include variations in viewpoint and resolution, occlusions, sensor noise, and partial or incomplete observations. The goal of this study is to estimate the sensor-to-map relative pose, even in the environments with these mentioned practical limitations.

Recent works have focused on a variety of solutions to the pose estimation between the RGBD sensor and point clouds maps, which can be broadly categorized into the following two aspects: traditional geometric and feature design solutions and learning-based approaches. Traditional methods typically reply on the hand-crafted features such as Scale-Invariant Feature Transform (SIFT) (Lowe, 1999), Speeded Up Robust Features (SURF) (Bay et al., 2006), Orientated Fast and Rotated Brief (ORB) features (Rublee et al., 2011). The designed features are often followed by the matching solution such as such as Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981), Normal Distribution Transform (NDT) (Biber, 2003), and Iterative Closest Point (ICP) (Segal et al., 2010). The traditional solutions can often achieve standard accuracy; however, the performance is unstable and often sensitive to the quality of the initial pose estimation and the initial overlap, noise data, and the point clouds density.

With the advantage of deep learning, learning-based solutions have gained significant traction, with the advantages of robustness to varied viewpoint, noised raw data, and occlusion of measurements. Deep networks encode the multi-modal features and match the heterogeneous data within a unified representation space. For example, recent publications such as CorrI2P (Ren et al., 2023), CoFiI2P (Kang et al., 2023), and 2D3D-MatchNet (Feng et al., 2019) introduce matching framework of image to point clouds by designing multi-modal encoders. Image encoder and point clouds encoder are designed to find the similarity, to find the corresponding key feature layer and calculate the relative pose.

Beyond finding feature similarity, some methods incorporate geometric constraints collected from RGBD sensors to improve the reliability of registration. For example, RGBD-Glue (Chen et al., 2024) proposes a feature combination framework fusing the visual feature descriptor and geometric feature descriptor, using an adaptive filter to estimate the alignment matrix. PointMBF (Yuan et al., 2023) proposes a bidirectional fusion network to find the correspondence estimation by calculating the photometric and geometric consistency losses. LLT (Wang et al., 2022) designs a geometry-aware visual feature extractor followed by a proposed local linear transformation module for the alignment. These works demonstrate improved performance, particularly under conditions of stable and dense point cloud coverage.

Despite of the current progresses, there are remaining limitations of the current registration frameworks, which can be discussed as following:

(1). Dependency on high-quality point clouds maps: Occlusions, sparsity, even inconsistency of the point clouds map can significantly reduce the accuracy of the RGBD point clouds registration.

(2). Sensitivity to initial positioning: Many methods reply on prior information or assumptions such as pose priors or high-overlap regions between the RGBD data and pre-built map, limiting the applicability.

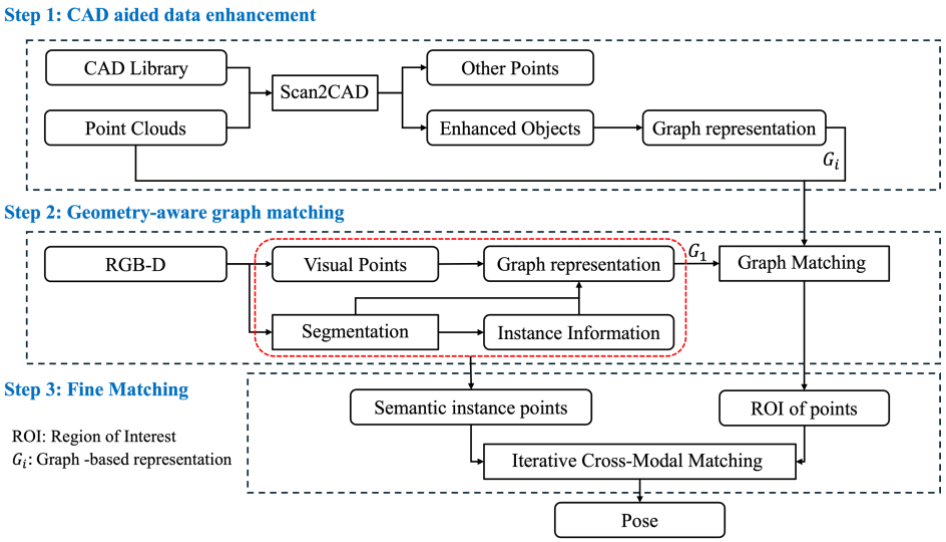

Figure 1. The proposed framework.

To address the mentioned challenges, this paper proposes a CAD model aided, multi-modal coarse-to-fine registration framework that estimate the relative pose between RGBD data and a pre-built point clouds map. The proposed framework includes two main stages: Firstly, with the aid of CAD models of functional objects (e.g., chairs and desks), data enhancement of the pre-built point clouds map is implemented. Secondly, a cross-modal coarse-to-fine registration framework is proposed to register the visual point clouds generated from RGBD sensor and the enhanced point clouds map, through a graph-based initial pose estimation and a refinement step. The proposed algorithm aims to reduce the dependency on dense point clouds quality and to improve the robustness in various of indoor environments, especially with occlusions, partial observations and viewpoint variations.

2. Methodology

2.1 Overview

The proposed RGBD-to-point cloud framework consists of three sequential steps. First, the indoor point clouds are enhanced by replacing the raw point cloud representations of instant functional objects using a pre-built CAD model library. This step enables accurate extraction of object centroids and semantic and instant information, facilitating improved geometric and semantic precision of the graph-based representation. Secondly, graph-based instance representation is implemented for the preparation of the initial pose estimation and the fine registration. For each extracted CAD model, the centroid of the bounding box and the semantic information is considered as the graph node with semantic labelling, while the pairwise Euclidean distance is calculated as the edge.

Through this process, instant objects within indoor environments can be represented as graph with topology information and semantic information. In addition, the input image is represented using another graph with the fusion of depth image and semantic information. A coarse matching strategy based on the constructed graph is then employed to identify regions of interest (ROIs), providing candidate environmental regions for further alignment. Finally, fine registration is performed by aligning the generated visual-semantic point clouds with the original indoor point clouds to achieve accurate spatial correspondence. An overview of the proposed framework is illustrated in Figure 1.

2.2 Multi-modal data enhancement

In this work, a data enhancement is firstly implemented to augment the data quality of indoor environment using Scan2CAD solution. By replacing the CAD models, centre of the object and the distance between objects can be easily extracted. To make this work easy to understand, a brief introduction of Scan2CAD is introduced (Dahnert et al., n.d.).

The Scan2CAD aims to align the CAD models to targeted objects in point clouds maps, providing the relative pose and corresponding instance semantic information. The Scan2CAD pipeline consists of the following key stages: Firstly, a 3D object detection is implemented to identify the potential object-level instances in the point clouds map, using a bounding box. The bounding box provides both location information and geometry information, served as initial candidates for the alignment. Secondly, a correspondence prediction network is employed, learning the point-level correspondence between the geometry of the detected candidates, and the corresponding CAD models. Finally, the alignment prediction is implemented to estimate a rigid transformation, which can align the CAD model to the determined and detected instance points. To demonstrate the results we had, Figure 2 shows an example of the result on the Scan2CAD model implemented on the ScanNet Dataset (Dahnert et al., n.d.; Dai et al., n.d.). For a better visualization, only replaced objects are visualized in the Figure 3, with random colors.

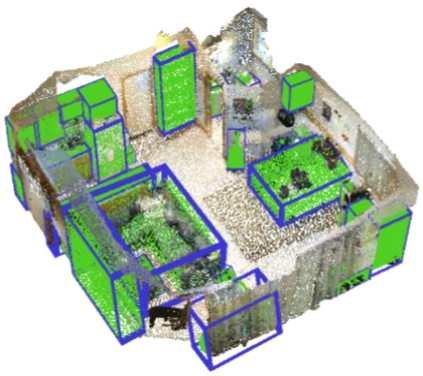

Figure 2. Example of the alignment results via Scan2CAD. Points sampled from CAD models are in green, while other points are coloured using RGB information. Blue boxes show the orientation and position of the CAD model.

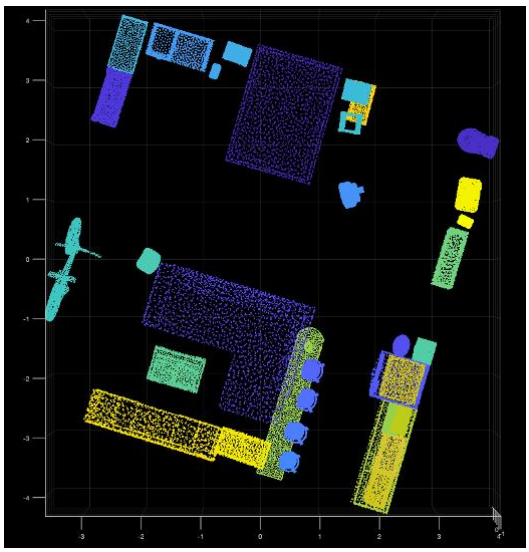

Figure 3. Visualization of the instance information from z-direction, while the color is randomly set.

$$Y = \frac{(u-c_y)\cdot Z}{f_y} \quad (2)$$

where $K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ is the intrinsic calibration matrix, while Z is the depth information provided by the depth camera $I = (u, v)$. The result of this process is dense point clouds within the range of view. An example of the generated colored point clouds can be found in Figure 4 (a).

Here, we assume the label information is aligned with the color frame. To implement the graph-based representation, a class-wise Euclidean clustering algorithm based on spatial proximity is implemented. The goal of this step is to extract the instance label and segment multiple objects with same class types. Given the 3D visual points generated from RGBD with labelling for each point, the points are firstly organized using a KD-tree structure to achieve the nearest-neighbour searching. The clustering process iteratively expands the local regions by grouping the points whose distances are below a pre-defined spatial threshold. Starting from a seed point, all neighbouring points are added to the same groups. Once the local neighbourhood is over, the algorithm continues to the next unvisited point and starts a new group. This cluster step provides a set of instance labels, supporting the graph-based representation. An example is shown in Figure 4 (b).

2.3.2 Semantic graph construction

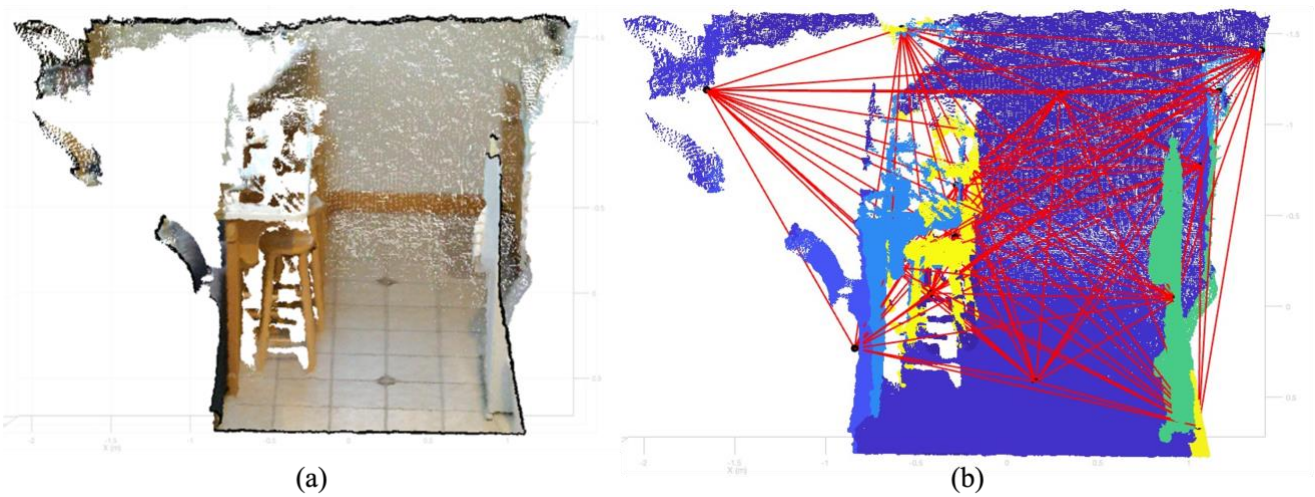

Figure 4. Example of the visual point clouds generated using RGBD (a), and the graph-based representation (b) Red lines demonstrate the edge between each center point. (Dataset: ScanNet, scene0000_01, Image 001700

2.3 Graph-based representation

This section introduces the detailed steps of object instances' graph-based representation. Following the Scan2CAD implementation, CAD models are aligned to the point clouds, and are sampled to dense point clouds. Visual point clouds and sampled point clouds are used for the semantic graph generation. Details are introduced in this following section.

2.3.1 Visual point clouds generation and graph-based representation: Firstly, the RGB-D image will be converted to point clouds using the intrinsic parameters and extrinsic calibration between the depth camera and the color camera. The process can be formulated using the following equations,

$$X = \frac{(u-c_x)\cdot Z}{f_x} \quad (1)$$

To facilitate robust and efficient coarse registration between the RGBD sensor and the pre-built map, a semantic graph $G = < V, E, A >$ is constructed to abstract key functional objects layout information from the observed scene. In this representation:

- V_{scan} and V_{map} denote the set of the nodes of the detected semantic object instance in the RGBD frame and the point clouds map, respectively.
- E_{scan} and E_{map} represents the set of edges encoding the topological and relationships between these functional objects.
- A demonstrates the semantic instance information which can be associated with each node.

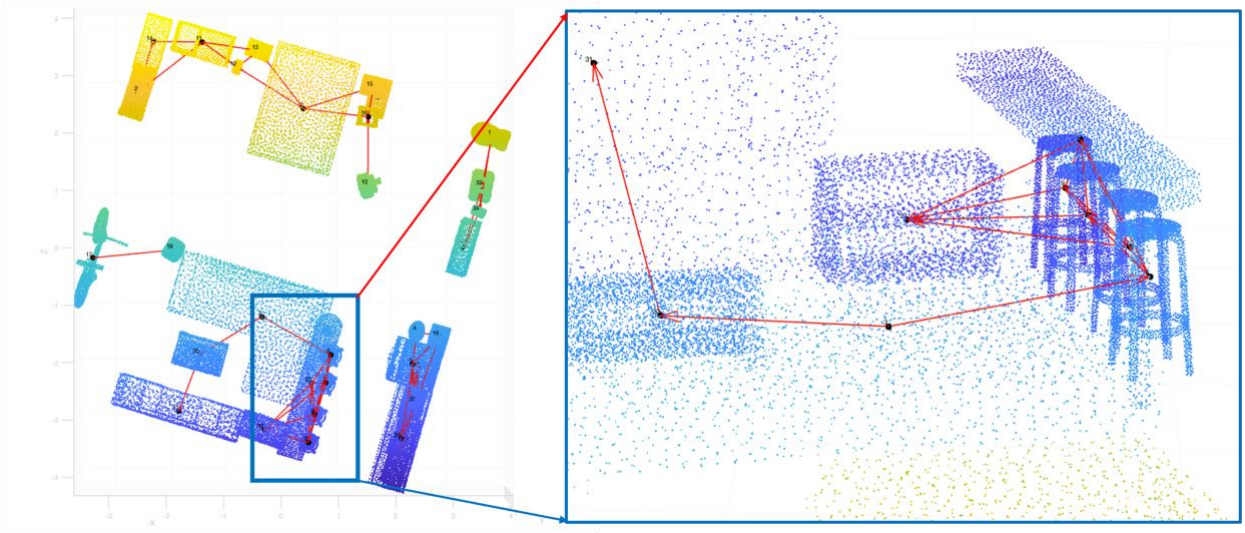

Figure 5. Example of the generated graph using the enhanced point clouds. After removing the unrelative edges using a pre-defined threshold δ, red lines are the edges, while point clouds are colored using Y value.

For each detected instant I_i, we represent it using the centre of the object $(\overline{x_i}, \overline{y_i}, \overline{z_i})$ for the sampled point clouds (X, Y, Z), composing the 3D node V_i, as following,

$$V_i = (\overline{x_i}, \overline{y_i}, \overline{z_i}) \tag{3}$$

To model the spatial relationship between instances, edges are calculated using the directional edge encoding, as following:

$$\mathbf{d}_{ij} = \frac{\mathbf{c}_j - \mathbf{c}_i}{\|\mathbf{c}_j - \mathbf{c}_i\|_2}, \text{ if } \|\mathbf{c}_j - \mathbf{c}_i\|_2 \leq \tau \tag{4}$$

where direction vectors $\mathbf{c}_k = \frac{1}{|\mathcal{I}_k|}\sum_{i \in \mathcal{I}_k} \mathbf{x}_i$, $\mathcal{I}_k$ is the set of point indices belonging to the instance k, and $\mathbf{c}_k$.

The graph-based representation can improve the robustness of the viewpoint changes and make it resilient to minor occlusion, as it abstracts the topological relationship and semantic information to extract higher -level scene structure. The visualization of the $G = < V_{map}, E_{map}, A >$ can be found in Figure 5 (a), while the blue lines demonstrate the edges connected between each two objects. Considering the spatial relationship, only center within a pre-defined range will be kept as the edges aiming to improve the efficiency of the whole framework.

2.4 Geometry-aware Coarse-level Graph matching

To determine the approximate initial position of the RGBD sensor within the pre-built point clouds map, a coarse-level graph matching is implemented using the generated semantic graphs. Specifically, the problem of the estimating the Region of Interest (ROI) is formulated as a graph similarity matching task between the generated two graphs: one generated from the visual point clouds generated from RGBD sensor, and another one generated from the entire point clouds map. This approach considered both semantic attributes and topological structure to estimate the initial pose of the RGBD sensor, serving as an initialization step for the next level of fine registration.

The objective is to identify the subgraph $G^* \in G_2$ that matches the G_1 considering both node similarity and edges distribution. The matching problem can be formulated as the following equation:

$$G^* = \arg \max_{G_i \in G_2} \text{Sim}\,(G_1, G_i) \tag{5}$$

Here, the Sim demonstrate the similarity score between two graphs, which can be calculated as a weighted combination of node similarity and edge similarity, using the following equation:

$$\text{Sim}\,(G_1, G_i) = \alpha \cdot S_V + \beta \cdot S_E \tag{6}$$

where α and β are normalization weights, balancing the influence of the nodes and edges terms.

The node similarity term S_V and edge similarity term S_E are calculated using the following equations:

$$S_V = \frac{1}{|V_1|}\sum_{v \in \text{scan}} \max_{v' \in V_i} \delta_L(v, v') \tag{7}$$

where $\delta_L(v, v') = 1$ if the semantic labels can be matched, and 0 otherwise.

$$\text{Sim}_{\text{dir}}\left(e_{ij}, e_{kl}\right) = \mathbf{d}_{ij}^{\top}\mathbf{d}_{kl} \tag{8}$$

where sim_{dir} aggregates directional cosine similarities over corresponding edges. An example of the matched nodes and edges are shown in Figure 6, while red lines demonstrate the matched edges and the corresponding nodes.

Following the geometry-aware graph matching process, the ROI is detected based on the similarity. Assuming $P = \{P_1, P_2 \ldots P_n\}$ and $Q = \{Q_1, Q_2 \ldots Q_n\}$ are the 3D centroids of the nodes of RGBD scan and the corresponding centroids detected in the global map. The goal is to estimate the rigid transformation matrix $T = [R, t] \in SE(3)$, consisting of the rotation matrix R and a translational matrix t. The process can be formulated in the following equation:

$$\mathbf{T}_{\text{init}} = \arg \min_{\mathbf{R},\mathbf{t}} \sum_{k=1}^{n} \|\mathbf{R}\mathbf{p}_k + \mathbf{t} - \mathbf{q}_k\|^2 \tag{9}$$

The optimal solution is calculated using pointset P and Q. The result will be served as the initial position of the pose estimation.

https://doi.org/10.5194/isprs-archives-XLVIII-1-W5-2025-1-2025

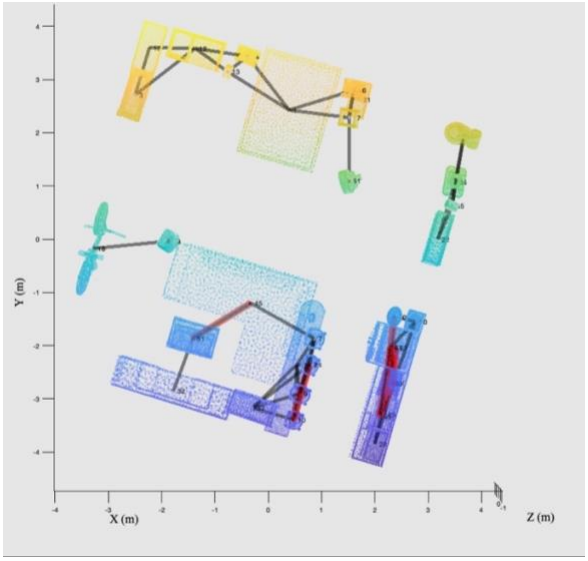

Figure 6. An example of graph matching. Black lines demonstrate the graph of the indoor map, while the red lines show the matched edges and nodes.

Here, only translational initial pose is used for the refinement$T_o = [0; t_0]$. The rotational initial value is set to zero. The final pose will be obtained once the error is less than a pre-defined threshold, enabling RGBD sensor's accurate pose in the point clouds map.

3. Experimental results

3.1 Experiment details

In this work, the process of data enhancement is implemented using Scan2CAD framework, and the data from ShapeNet and ModelNet. Indoor scene 001 from ScanNet is selected as experimental area, while three RGBD data are used for the registration as raw data. Based on the assumption of well - calibration, extrinsic and intrinsic calibration files are used for visual point clouds generation. Pre-defined threshold used in this paper are listed in the Table 1.

Meaning	Parameters/Value	
Edge removal	δ	1.5 m
ROI	γ	3 m
Residual threshold	ε	10-3e

Table 1. Pre-defined parameters and threshold

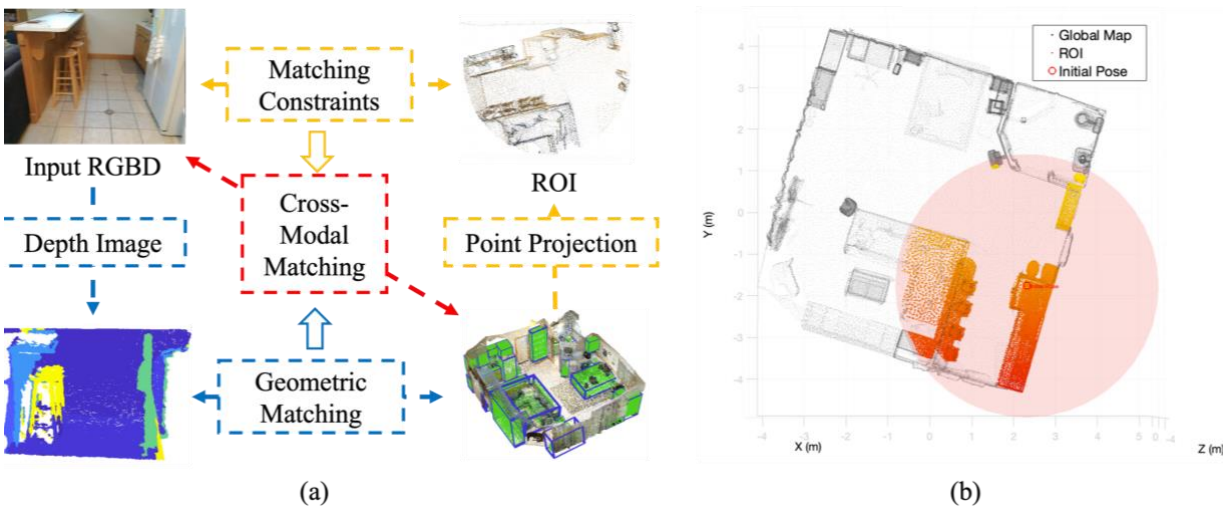

Figure 7. (a) Demonstration of Cross-Modal Matching. (b) Example of the ROI detection.

2.5 Iterative cross-modal matching for pose refinement

After obtaining the initial pose estimation $\mathbf{T}_{\text{init}}$ via coarse-level sgraph alignment, an iterative cross-modal fine matching process is introduced to refine the pose matrix, as demonstrated in Figure 7 (a). The ROI is selected based on the position of the initial position, with the radius of pre-defined threshold γ, as shown in Figure 7 (b). To implement the cross-modal matching, both 3D geometric and visual semantic correspondence is constructed. Via a colored information aided G-ICP, the iterative registration will be stop until the regsistration error meets the pre-defined threshold (Korn et al., 2014). By using the G-ICP, with the aid of consistency of colored information, the objective function of pose refinement can be formulated as,

$$\Delta\mathbf{T} = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{k=1}^{|\mathcal{C}|} w_k \cdot \|\Delta\mathbf{R}\mathbf{P}_k + \Delta\mathbf{t} - \mathbf{Q}_k\|^2 \quad (10)$$

The final transformation is iteratively updated, using the following equation:

$$\mathbf{T}_{t+1} = \Delta\mathbf{T} \cdot \mathbf{T}_0 \quad (11)$$

3.2 Experimental results and performance evaluation

Figure 9 illustrates an example of the RGBD-to-point cloud registration, where the colored RGBD visual points are aligned within the Region of Interest (ROI) successfully, with a translational registration error at 0.095 m. To have a quantitatively evaluation, the error calculation includes both the mean and the maximum translational errors, providing a comprehensive assessment of accuracy. As shown in Table 2, the results demonstrate that the proposed framework achieves reliable and precise alignment performance.

Value	Translational error
Average	0.12 m
Max error	0.16 m

Table 2. Translational errors.

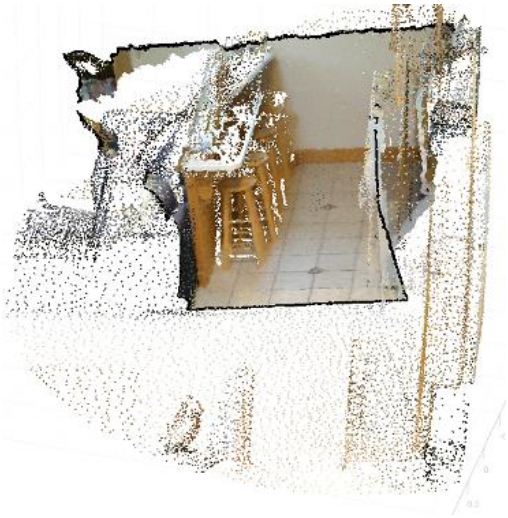

Figure 9. Visualization of the registration results (Image 0017 and Indoor scene 001).

4. Conclusion

This paper presents a CAD-aided cross-modal pose estimation framework for RGBD sensors operating in pre-built indoor point clouds map. The proposed method firstly adopts the Scan2CAD framework as a data enhancement step, using the CAD model to represent the spatial information and semantic instances. Secondly, the framework proposes a coarse-to-fine strategy, beginning with geometry-aware semantic graph matching to determine the ROI and estimate the initial pose, enabling the efficient matching across large-scale indoor maps. Following the coarse matching, an iterative cross-modal fine matching algorithm that fuses both visual features and geometric features to establish reliable correspondences, via a G-ICP. The final output is the estimated pose of the RGBD sensor in the point clouds map, supporting the applications such as robotic navigation and positioning. Experiments are implemented on opensource dataset, demonstrating that the proposed framework can accurately estimate the pose of RGBD sensor in point cloud maps.

The limitation of proposed approach can be discussed in the following two aspects: (1). The refine mapping still rely on hand-crafted feature description and heuristic matching pipelines, which may limit the ability of the highly dynamic environments. (2). An assumption of this paper is that the semantic graph matching relies on the topological relationship between clear objects. Clear objects within an image are required for the graph matching. The initial pose estimation will fall if there is no clear object in the RGBD data.

Aiming the limitation, future works will focus on developing an end-to-end deep learning-based pose estimation framework that directly infers the transformation between the RGBD input and the global point clouds map. By replacing the modular steps with fully trainable networks, the system can potentially lean richer cross-modal correspondences. In addition, the domain adaptation techniques for cross-modal recognition will be developed to further enhance the performance.

Acknowledgements

This research has been supported by funding of Mitacs program and Micro Engineering Tech Inc.

References

Asl Sabbaghian Hokmabadi, I., Ai, M., El-Sheimy, N., 2023. Shaped-Based Tightly Coupled IMU/Camera Object-Level SLAM. Sensors 23. https://doi.org/10.3390/s23187958

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded up robust features. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 3951 LNCS, 404–417. https://doi.org/10.1007/11744023_32

Biber, P., 2003. The Normal Distributions Transform: A New Approach to Laser Scan Matching. IEEE Int. Conf. Intell. Robot. Syst. 3, 2743–2748. https://doi.org/10.1109/iros.2003.1249285

Chen, C., Jia, X., Zheng, Y., Qu, Y., 2024. RGBD-Glue: General Feature Combination for Robust RGB-D Point Cloud Registration.

Dahnert, M., Dai, A., Chang, A.X., Nießner, M., n.d. Scan2CAD: Learning CAD Model Alignment in RGB-D Scans.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., n.d. ScanNet : Richly-annotated 3D Reconstructions of Indoor Scenes.

Feng, M., Hu, S., Ang, M.H., Lee, G.H., 2019. 2D3D-matchnet: Learning to match keypoints across 2d image and 3d point cloud. Proc. - IEEE Int. Conf. Robot. Autom. 2019-May, 4790–4796. https://doi.org/10.1109/ICRA.2019.8794415

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM 24, 381–395. https://doi.org/10.1145/358669.358692

Kang, S., Liao, Y., Li, J., Liang, F., Li, Y., Li, F., Dong, Z., Yang, B., 2023. CoFiI2P: Coarse-to-Fine Correspondences for Image-to-Point Cloud Registration 1–8.

Korn, M., Holzkothen, M., Pauli, J., 2014. Color supported generalized-ICP. VISAPP 2014 - Proc. 9th Int. Conf. Comput. Vis. Theory Appl. 3, 592–599. https://doi.org/10.5220/0004692805920599

Lowe, D.G., 1999. Object recognition from local scale-invariant features. Proc. IEEE Int. Conf. Comput. Vis. 2, 1150–1157. https://doi.org/10.1109/iccv.1999.790410

Ren, S., Zeng, Y., Hou, J., Chen, X., 2023. CorrI2P: Deep Image-to-Point Cloud Registration via Dense Correspondence. IEEE Trans. Circuits Syst. Video Technol. 33, 1198–1208. https://doi.org/10.1109/TCSVT.2022.3208859

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. Proc. IEEE Int. Conf. Comput. Vis. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

Segal, A. V, Haehnel, D., Thrun, S., 2010. Generalized-ICP. Robot. Sci. Syst. 5, 161–168.

Wang, Z., Huo, X., Chen, Z., Zhang, J., Sheng, L., Xu, D., 2022. Improving RGB-D Point Cloud Registration by Learning Multi-scale Local Linear Transformation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 13692 LNCS, 175–191. https://doi.org/10.1007/978-3-031-19824-3_11

Yuan, M., Fu, K., Li, Z., Meng, Y., Wang, M., 2023. PointMBF: A Multi-scale Bidirectional Fusion Network for Unsupervised RGB-D Point Cloud Registration. Proc. IEEE Int. Conf. Comput. Vis. 17648–17659. https://doi.org/10.1109/ICCV51070.2023.01622