Image-to-Point Cloud Registration using Camera Motion Generation and Monocular Depth Estimation

Masoud Kamali [1], Behnam Atazadeh [1], Abbas Rajabifard [1], Yiqun Chen [1]

[1] The Centre for Spatial Data Infrastructures and Land Administration, Department of Infrastructure Engineering, The University of Melbourne, Parkville, 3010, Victoria, Australia
(m.kamali, behnam.atazadeh, abbas.r, yiqun.c) @unimelb.edu.au

Keywords: GSW 2025, Image-to-point cloud registration, Monocular depth estimation, Camera motion generation

Abstract

Image-to-point cloud (I2P) registration plays a vital role in applications requiring accurate spatial alignment between 2D images and their corresponding 3D point clouds. Traditional I2P methods often require extensive training to generalise across diverse environments and rely on intrinsic camera parameters for accurate metric depth estimation, limiting their effectiveness in complex or unseen scenarios. To address these challenges, this study introduces a novel approach that leverages Camera Motion Generation (CMG) and Monocular Depth Estimation (MDE) for I2P registration task. CMG simulates camera movements in the up, down, left, and right directions, enabling the generation of novel viewpoints of the scene. MDE is applied to each frame to generate point clouds, which are subsequently registered using multi-way registration. The final registered point cloud is then aligned with the scene point cloud through the Iterative Closest Point (ICP) algorithm, ensuring precise spatial alignment. The proposed method eliminates the need for training or reliance on intrinsic camera parameters, making it robust for diverse and unseen environments. We evaluated the proposed approach through extensive experiments using the Root Mean Square Error (RMSE) to measure registration accuracy between the generated and ground truth point clouds. The results indicate that our method achieves competitive RMSE values across various scenarios, validating its effectiveness in enhancing I2P registration accuracy and adaptability.

1. Introduction

Image-to-point cloud (I2P) registration involves determining the rigid transformation, encompassing both rotation and translation, that aligns a 3D point cloud with its corresponding 2D image (Kang et al., 2023). I2P registration has numerous applications in fields such as robotics and augmented reality (Bai et al., 2024). To establish accurate pixel-to-point correspondences in I2P registration, depth information is essential. Depth information allows each pixel in the 2D image to be associated with a corresponding 3D point in the point cloud, facilitating precise spatial alignment. This information is often obtained from sensors such as LiDAR, which provide direct measurements of the distance between the sensor and the observed scene. While LiDAR offers precise and direct depth measurements, they have several limitations that affect their practicality in certain applications. LiDAR systems are often expensive, and power-intensive, making them less suitable for lightweight, low-cost, or portable scenarios.

Monocular depth estimation (MDE) is the process of estimating depth information from a single image to predict the distance from the camera to each pixel within the scene (Ke et al., 2024). Unlike stereo cameras or LiDAR systems, MDE does not require multiple viewpoints or specialised sensors, making it a more accessible and cost-effective solution for depth estimation. Despite the challenges associated with MDE, particularly regarding accuracy in complex or unseen environments, recent advancements in deep learning, coupled with the availability of large-scale datasets, have improved its performance in zero-shot depth estimation. Consequently, MDE has gained increased relevance in I2P registration (Wang et al., 2023). Depth estimation can be classified into two categories: (1) relative depth estimation and (2) metric depth estimation. Relative depth estimation techniques assess the depth of objects within an image in relation to one another, offering a spatial understanding of the scene's layout. While this approach enhances scene perception, it is insufficient for I2P registration, which requires precise, absolute depth measurements to accurately align images with 3D point clouds. In contrast, metric depth estimation provides depth values in real-world units, such as metres, representing the actual distance from the camera to each object, making it suitable for registration tasks.

Point clouds and multi-view images provide a detailed representation of scenes by combining depth information with visual features in 3D environment. However, when using single images, the visual context is often insufficient, presenting significant challenges for accurate I2P registration. Single images capture the scene from a single viewpoint, which restricts the ability to infer depth and structure for aligning 2D image data to 3D point clouds. Additionally, single viewpoint is inadequate for capturing geometric relationships and occlusions within the scene. Video generation from a single image can produce multiple frames that simulate different viewpoints and movements, thereby enriching the visual information extracted from a single image. Image-to-video (I2V) generation allows for a rich representation of the environment by extrapolating additional visual information that is not present in the original image (Gupta et al., 2022). By synthesising multiple frames, this technique offers additional visual cues that can help in tasks such as scene interpretation and 3D reconstruction (Voleti et al., 2024).

In video sequences, motion can be classified into two types: (1) local motion and (2) global motion (Wang et al., 2024b). Local motion refers to the movement of objects within a scene while the camera remains fixed. This type of motion changes the relative positions, interactions, and dynamics of the objects without altering the viewpoint. This is essential for simulating dynamic objects to enhance the overall realism of the video. Global motion refers to the movement of the camera, which results in a shift in the view of the entire scene. As the camera moves, the corresponding shift in viewpoint alters the spatial representation of objects, affecting both their position and scale within the scene. This type of motion enhances the visual

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-453-2025

perspective and plays a crucial role in improving the understanding of spatial relationships within the scene. In this study, we refer to this process as camera motion generation (CMG), which involves creating simulated or controlled camera movements to obtain multiple viewpoints of the scene.

This study proposes a novel method using CMG to address the limitations of depth estimation and visual information in current I2P registration tasks that rely solely on single images. Camera motions are simulated using the MotionCtrl (Wang et al., 2024b) framework, which generates multiple frames from a single image by simulating various camera movements. From the generated video, appropriate frames are extracted for each motion direction to ensure the highest quality and consistency. For depth estimation, the Depth Anything V2 model (Yang et al., 2024b) is employed to perform monocular metric depth estimation on both the original image and the selected frames from each motion. This depth information is then used to generate point clouds for each image, providing a 3D spatial representation of the scene. These point clouds are then aligned using multiway registration (Choi et al., 2015), ensuring that they are consistently positioned within a global coordinate system. In this process, pose graph optimisation is applied to align the generated point clouds by calculating transformations between overlapping geometries. Following this, an initial alignment is conducted to reduce the spatial disparity between the registered point clouds and the ground truth point cloud. Finally, Iterative Closest Point (ICP) registration is applied to further refine the alignment. To assess the effectiveness of the proposed method, we evaluate the accuracy of the I2P registration by comparing the registered point clouds with the ground truth point cloud. The assessment is primarily conducted using the Root Mean Square Error (RMSE), which measures the average distance between corresponding points in the registered and ground truth point cloud. In summary, the key contributions of this study are as follows:

1. To the best of our knowledge, this is the first study to utilise CMG for the I2P registration task.
2. A method is designed to enable I2P registration in unseen environments without requiring any training.
3. The reliance on intrinsic camera parameters for metric depth estimation is eliminated by employing zero-shot MDE models.
4. Experiments are conducted to demonstrate the feasibility of the proposed approach for I2P registration.

The remainder of this paper is organised as follows. The literature review is presented in Section 2. In Section 3, we detail the proposed methodology. Section 4 provides the results and discussions of the findings. Finally, the conclusions and potential future works are outlined in Section 5.

2. Literature Review

This section reviews the state-of-the-art (SOTA) algorithms for I2P registration task. Recent advancements in MDE methods are then examined. Finally, CMG models are explored, and their relevance to the current study is discussed.

2.1 Image-to-Point Cloud Registration

To register image into the point clouds, most existing approaches utilise high-quality training datasets (Kang et al., 2023, Li and Lee, 2021, Jeon and Seo, 2022) and depth information (Campbell et al., 2020, Liu et al., 2020, Wang et al., 2021). Besides, some approaches require camera intrinsic parameters for 2D-to-3D transformation (Sheng et al., 2024). CoFiI2P (Kang et al., 2023) introduced a registration network that progressively aligns images and point clouds by extracting multi-level correspondences using transformers, incorporating both self-attention and cross-attention mechanisms to enhance the robustness and accuracy of image-to-point cloud registration in autonomous systems. DeepI2P (Li and Lee, 2021) presented an approach for I2P registration, transforming the registration problem into a two-stage classification and optimisation framework. This approach predicted the rigid transformation by estimating whether points in the point cloud project within or outside the camera frustum. EFGHNet (Jeon and Seo, 2022) introduced an I2P registration method for outdoor environments, utilising a two-phase process—virtual alignment and compare-and-match to estimate the transformation between an image and a pre-collected point cloud for both urban and off-road settings. CorrI2P (Ren et al., 2022) developed a feature-based dense correspondence framework for I2P registration, leveraging a two-branch neural network with a symmetric overlapping region detector to extract dense 2D-3D correspondences and estimate the camera pose. CFI2P (Yao et al., 2024) proposed a coarse-to-fine correspondence learning framework for image-to-point cloud registration, emphasising quantity-aware correspondences between point sets and pixel patches to improve matching accuracy. All these learning-based approaches employed extensive point cloud data and corresponding images datasets for network training. Besides, they were specifically designed for outdoor environments, limiting their ability to generalise to indoor settings. In this study, a pre-trained monocular depth estimator is employed to eliminate the need for training on the I2P registration, with a primary focus on indoor environments.

2.2 Monocular Depth Estimation

The primary challenge in MDE lies in estimating depth information from a single 2D image, which is fundamentally ambiguous due to the absence of 3D spatial data. Several studies have focused on both relative (Ranftl et al., 2020, Yao et al., 2024, Ke et al., 2024) and metric (Bhat et al., 2023, Yin et al., 2023, Yang et al., 2024a) MDE. Marigold (Ke et al., 2024) introduced a diffusion-based model for relative MDE, utilising pre-trained image stable diffusion and fine-tuning them with synthetic data to achieve SOTA performance on natural images, even in zero-shot scenarios using only synthetic RGB-D data. ZoeDepth (Bhat et al., 2023) presented a two-stage MDE that integrates relative depth pre-training with metric depth fine-tuning. It employed an encoder-decoder architecture for relative depth estimation, then adds domain-specific heads for metric depth using a lightweight metric bins module. Depth Anything (Yang et al., 2024a) proposed a relative and metric MDE that scales up by leveraging large-scale unlabelled datasets, combining self-supervised learning and rich semantic priors from pre-trained encoders to achieve strong zero-shot generalisation across unseen scenes. Depth Anything V2 (Yang et al., 2024b) enhances the Depth Anything model by replacing labelled real images with synthetic data, scaling up the dataset with pseudo-labelled real images (62M), and utilising a teacher-student model framework for training. It provides more robust predictions for complex scenes, better generalisation through metric depth fine-tuning, and faster inference speeds. According to the accuracy of SOTA pre-trained models, Depth Anything V2 is utilised for zero-shot metric MDE in this research.

2.3 Camera Motion Control in Image-to-Video Generation

I2V generation refers to the process of generating consistent video sequences from static images. Traditionally, I2V methods have utilised approaches such as Generative Adversarial

Networks (Wang et al., 2020, Tulyakov et al., 2018) and Variational Autoencoders (Wang et al., 2022, Xu et al., 2023) However, with the exceptional performance of diffusion models in image generation tasks, I2V research has shifted towards leveraging diffusion models for advanced video generation (Blattmann et al., 2023, Guo et al., 2023). CMG in I2V models is essential for producing realistic and dynamic video sequences, as it effectively simulates natural movements. MotionLoRA is a lightweight fine-tuning technique integrated into AnimateDiff (Guo et al., 2023), which allows a pre-trained motion module to efficiently adapt to new motion patterns. This method leverages LoRA layers in the self-attention components of the motion module, enabling personalisation of motion effects with limited reference videos. VideoComposer (Wang et al., 2024a) incorporated motion control by utilising motion vectors extracted from compressed videos as temporal control signals. These vectors guide the synthesis of inter-frame dynamics, ensuring temporal consistency and enabling the generation of smooth, controlled video sequences. MotionCtrl (Wang et al., 2024b) utilised a multi-step training strategy to control both camera and object motion effectively by training on distinct datasets tailored to specific motion control needs. The Camera Motion Control Module (CMCM) is trained using the Realestate10K (Zhou et al.). While the dataset has limitations in scene diversity, it provides precise annotations of camera poses that enhance the quality of training. Challenges such as the lack of captions for text-to-video models are addressed by integrating Blip2 (Li et al., 2023) to generate the necessary captions for video clips in Realestate10K. Given the superior performance of MotionCtrl compared to MotionLoRA and VideoComposer (Wang et al., 2024a), CMCM of MotionCtrl was selected for CMG in this study.

3. Method

Figure 1 illustrates the research methodology in this study. The process begins with a single input image, where camera motion is simulated to generate a sequence of frames for each movement. Afterward, a metric depth estimator is applied individually to both the original and extracted frames. Point clouds are generated from the depth information of each frame and are then aligned through multiway registration. An initial alignment is performed to refine the transformation of the geometries. Finally, the RMSE is calculated using the ICP algorithm to assess the accuracy of the alignment.

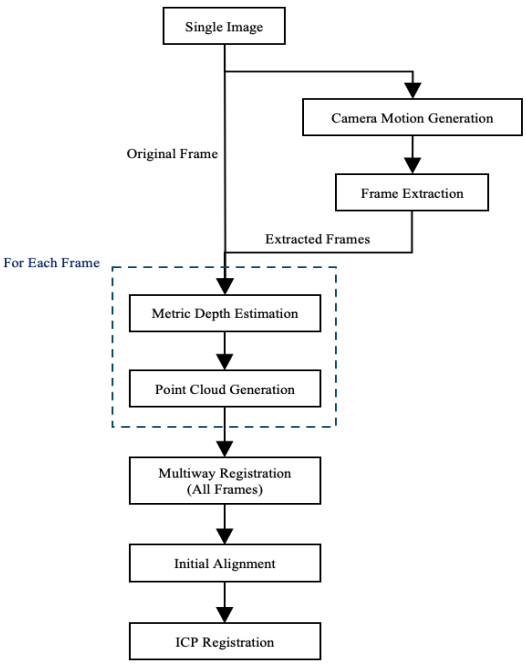

Figure 1. Research methodology

3.1 Camera Motion Generation

CMG refers to the simulation of dynamic camera movements to generate a sequence of frames from a single image. This technique enables the creation of video sequences by controlling the virtual camera’s motion, offering advantages in scenarios where specific camera movements are necessary to interpret or analyse scenes in ways that cannot be achieved with the original image. In this study, MotionCtrl is utilised to generate video from a single image by controlling the camera's movement. The CMCM module takes a series of camera poses, denoted as $RT = \{RT0, RT1, ..., RTL-1\}$, as input. In this module, the camera pose is described using a 3×3 rotation matrix and a 3×1 translation matrix. According to the capabilities of MotionCtrl, four main camera motions, including up, down, right, and left, are simulated to generate one-second videos for each movement. The seed value is set to 1230, and the frame rate is configured at 10 frames per second. Figure 2 depicts the frames from a one-second video generated for each main camera motions.

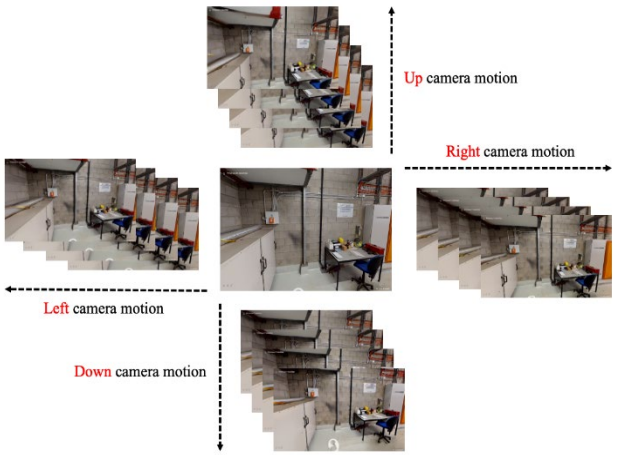

Figure 2. Camera motion generation for four main movements

3.2 Frame Extraction

In the frame extraction process, an initial set of 30 frames was extracted from each one-second video generated by the camera motion simulation. However, after careful evaluation, issues such as geometric distortions and blurriness in the scenes were observed from a certain frame onwards. To address the problem, the last consistent frame was selected for each motion. This approach ensured that the selected frame preserved both visual clarity and consistency. Figure 3 shows the frames extracted from a generated video with MotionCtrl. As shown in the figure, some frames exhibit issues such as changes in object geometry and blurriness, particularly in later frames.

3.3 Metric Depth Estimation

The process of metric depth estimation from a single image involves using computational models to infer the 3D structure of a scene from its 2D representation. Due to the absence of camera intrinsic parameters, it is essential to employ zero-shot monocular metric depth estimation models. These models are specifically designed to infer depth information from single images without requiring explicit calibration data. This study leverages Depth Anything V2 model to estimate depth information of single images in unseen scenarios. This model is

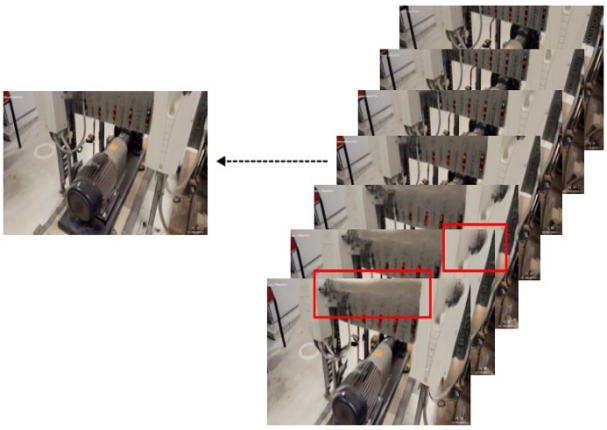

Figure 3. Frame extraction from the generated video (with "up" camera motion as the example)

designed to enhance monocular depth estimation by utilising a teacher-student learning approach. Initially, the model employs DINOv2-G (Oquab et al., 2023), which is trained exclusively on high-quality synthetic images known for their precision but limited diversity. To address the challenges of distribution shift, pseudo labels were generated for a diverse set of unlabelled real-world images. Finally, the student models were trained on these pseudo-labelled real images, enabling robust generalisation across varied environments. The pre-trained encoder was transferred for metric depth estimation. To enhance real-world applications like multi-view synthesis, the encoder was fine-tuned using the Hypersim (Roberts et al., 2021) for indoor environments and the Virtual KITTI (Cabon et al., 2020) for outdoor scenarios, ensuring robust performance in both domains of metric depth estimation. In this study, the Depth-Anything-V2-Large pre-trained model is employed as an encoder for metric depth estimation. Based on the characteristics of the captured indoor environment and the image dataset, the Hypersim dataset is used with a maximum depth threshold of 5 metres to estimate the depth information. As the intrinsic camera parameters are not available, the model utilises a default focal length of 470.4 for both the x and y axes. Figure 4 shows the pipeline to train Depth Anything V2 model.

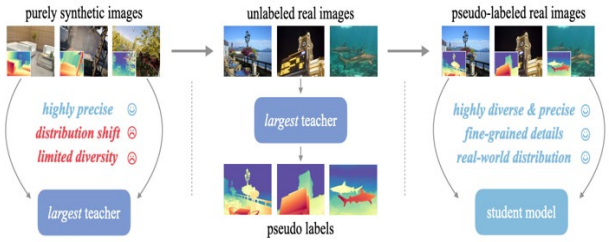

Figure 4. Depth Anything V2 training pipeline (Yang et al., 2024b)

3.4 Point Cloud Generation

To generate a point cloud from depth information, each 2D pixel in the image is transformed into a 3D point using its corresponding depth value. The 2D pixel coordinates are first normalised by subtracting the image centre and dividing by the default focal lengths of Depth Anything V2 model, resulting in normalised coordinates. These normalised coordinates are then multiplied by the depth value to obtain the 3D. Specifically, the 3D coordinates are computed as (Hartley and Zisserman, 2003):

$$(X, Y, Z) = \left(\frac{\left(x - \frac{width}{2}\right)}{f_x} \times z, \frac{\left(y - \frac{height}{2}\right)}{f_y} \times z, z \right), \quad (1)$$

where X, Y, Z = 3D coordinates
x, y = 2D pixel coordinates from the image
z = depth value for each pixel
f_x, f_y = the default focal length values (470.4 in Depth Anything V2 model)

The point cloud generated for each direction is stored in pcd format for multiway registration. Figure 5 illustrates the point cloud generation from the corresponding depth map.

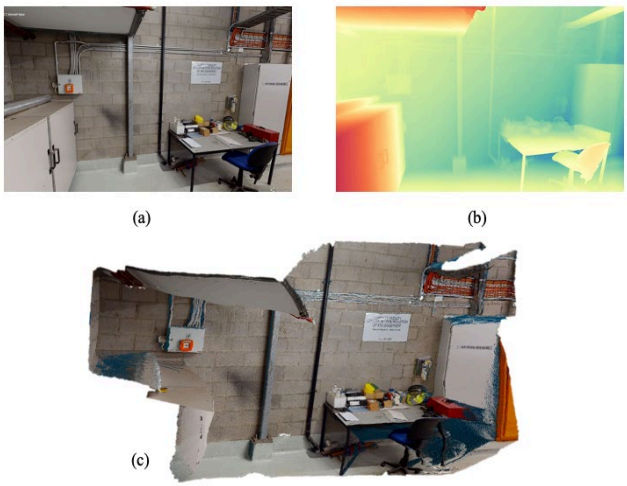

Figure 5. Point cloud generation from depth information. (a) RGB image; (b) Depth map; (c) Point cloud

3.5 Multiway Registration

Multiway registration is a technique used to align multiple geometries, such as point clouds, into a unified global space. This approach computes a set of rigid transformations, which ensures that the input geometries are correctly aligned after transformation. In this research, multiway registration is implemented using pose graph optimisation.

3.5.1 Pose Graph Optimisation

The pose graph consists of nodes, each representing a geometry, and edges, which define the relationships between geometries with overlapping regions. Each node is associated with a pose matrix, responsible for transforming the geometry into the global coordinate space. The global space is initialised by setting the pose of the first geometry. The transformations for the other geometries are derived based on pairwise registration between neighbouring nodes. Pairwise registration is performed using the point-to-plane ICP algorithm (Chen and Medioni, 1992), which minimises the alignment error by considering the surface normal of points. The objective function for point-to-plane ICP is formulated as:

$$E(T) = \sum_{(p,q) \in C} ((p - Tq).\overrightarrow{n_p})^2, \quad (2)$$

where $E(T)$ = object function that needs to be minimised.
p, q = the corresponding points from the source and target geometries
Tq = the point q from the source geometry after it has been transformed by the transformation matrix T

$\overrightarrow{n_p}$ = the normal vector of the surface at the point p in the target geometry

For each pair, the point-to-plane ICP algorithm is employed to calculate both coarse and fine transformations, which are subsequently integrated as edges into the pose graph. Odometry edges link neighbouring nodes that exhibit significant overlap, while loop closure edges connect non-adjacent nodes to ensure consistency across the entire global registration.
Once the pairwise registration step is complete, global optimisation is applied to refine the pose graph. This optimisation minimises the residuals from both the odometry and loop closure edges, while pruning false alignments and improving the accuracy of the transformation matrices. The optimised transformations are then applied to each geometry, ensuring that all point clouds are accurately aligned in the global coordinate space. In this study, the Open3D (Zhou et al., 2018) is utilised to implement multiway registration. This method ensures robust and precise alignment of geometries by leveraging both local and global geometric relationships.

3.6 Initial Alignment

Since the ground truth point cloud is a cropped subset of the complete environment and is positioned far from the multiway-registered point cloud in the global coordinate system, an initial coarse alignment is manually performed to roughly overlay the digital model onto the real-world reference before further processing. Without this preliminary step, the spatial distance between the ground truth and the registered point clouds would be too significant, hindering the effectiveness of ICP registration and the accurate calculation of the RMSE between the two point clouds.

3.7 ICP Registration

ICP registration is used to refine the alignment between the ground truth and the multiway registered point cloud. The ICP algorithm is an iterative method designed to minimise the difference between two point clouds by aligning them as closely as possible based on geometric correspondences. This algorithm transforms one point cloud (referred to as the "source") to minimize its distance from the other point cloud (the "target"). During each iteration, the transformation matrix was adjusted to minimise the alignment error, which is typically measured as the mean square distance between corresponding points in the two point clouds. Once the ICP registration was completed, RMSE was then calculated, providing a quantitative measure of the alignment accuracy. The software CloudCompare is used for ICP registration and RMSE calculation. Figure 6 shows the ICP registration results between two point clouds.

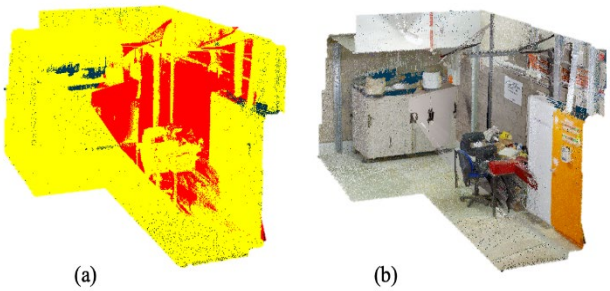

Figure 6. ICP registration results. (a) The yellow is the source point cloud and red is the target point cloud; (b) The colourised point clouds after ICP registration

4. Experiments and results

4.1 Experiment Settings

In this study, 2 NVIDIA L40 GPU with 48GB of VRAM, 16 vCPUs, and 241GB of system memory were utilised for CMG, MDE, and multiway registration tasks. The Ubuntu 20.04 has been used as operating system. Ground truth point cloud data was captured using a LEICA BLK360 laser scanner, while Agisoft Metashape and CloudCompare were employed for the point cloud processing.

4.2 Results and Discussion

In this section, we present and analyse the quantitative and qualitative outcomes of the I2P registration tasks.

4.2.1 Camera Motion Generation

For the CMG, one-minute videos demonstrating the four primary camera movements were generated using MotionCtrl. Figure 7 presents the CMG results for each direction. Each direction displays the generated results for "Scene 1" in the first row and "Scene 2" in the second row.

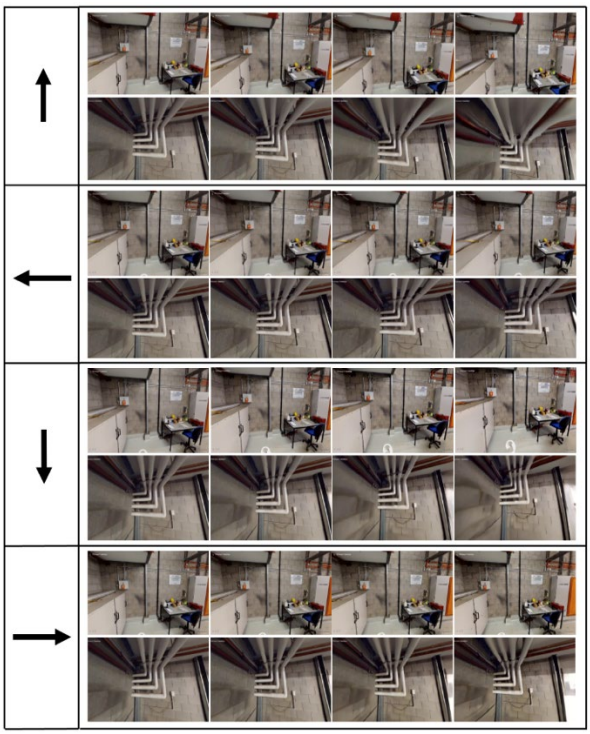

Figure 7. Video frames of CMG for four main movements

In both scenes, the extension of objects along the edges in each directional movement contributes to a detailed understanding of features within the scene and improves depth estimation. For example, in the rightward movement frames, the orange electrical conduit becomes more visible, allowing for clearer interpretation based on its colour and shape. This improves visibility assists in recognising the object and adds to the overall understanding of the scene. In the upward movement of Scene 2, the white pipes extend more clearly, providing a more detailed view of their structure. This outcome aligns with the ground truth data, indicating that the CMG method effectively captures the vertical layout. Similarly, the leftward movement in Scene 1 uncovers more of the white storage box, offering a clearer representation of its alignment within the scene. This visibility enhances depth

estimation by providing additional information about the object's spatial configuration. In the downward movement, the extension of the chair legs and the floor area offers a more detailed view, which can aid in the I2P registration process by providing additional points for alignment.

4.2.2 Multiway Registration

To evaluate the multiway registration of different camera motions, the RMSE was calculated for the combined point clouds generated by the MotionCtrl and Depth Anything V2 models. Figure 8 presents the RMSE results of multiway registration for Scene 1 using two voxel sizes (0.01 and 0.005 metres). The RMSE values indicate the effectiveness of the point-to-plane ICP algorithm in aligning the point clouds relative to one another for different movements. On the x-axis, arrows indicate the directions of the frames extracted from each motion. The circle represents the original image, while the “3-step” refers to the combination of three multiway registrations: (1) up, right, down; (2) up, left, down; and (3) the multiway registration of the first and second steps with the original image. The y-axis, measured in metres, shows the final RMSE after the multiway registration of camera motions and the original image.

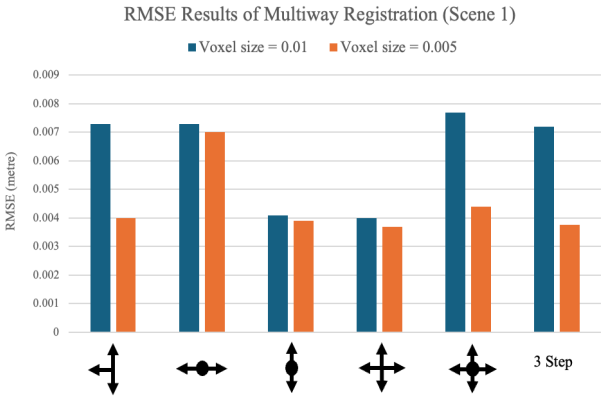

Figure 8. RMSE results of multiway registration for Scene 1

For the voxel size of 0.01 metres, the highest RMSE value was observed for the "Original image, Left, Up, Right, Down" combination with a value of 0.0077, indicating that combining multiple complex directional movements increases the overall registration error. When the voxel size was reduced to 0.005 metres, the overall RMSE values decreased, showing improved accuracy. The lowest RMSE value, 0.0037, was achieved by the "Left, Up, Right, Down" combination, indicating that finer voxel resolution significantly enhances registration accuracy by capturing more detailed alignments. In Scene 1, the multiway registration combinations that involved movements in the up and down directions exhibited lower RMSE values compared to those involving left and right movements. This is because the up and down movements primarily capture the walls, which have less variation in depth and fewer intricate features.

Figure 9 shows the RMSE results of multiway registration for Scene 2. For the voxel size of 0.01 metres, the highest RMSE value was observed in the "Original image, Left, Right" combination at 0.0089, while the lowest RMSE value for this voxel size was achieved by the "Left, Up, Right, Down" combination at 0.0080, suggesting that incorporating both vertical and horizontal movements can help reduce registration errors. When using the finer voxel size of 0.005 metres, the overall RMSE values decreased, indicating improved accuracy. The lowest RMSE was again found in the "Left, Up, Right, Down" combination with a value of 0.0044. In Scene 2, since the complexity of the scene is relatively consistent across both horizontal and vertical directions, the RMSE values are closely aligned across the different combinations.

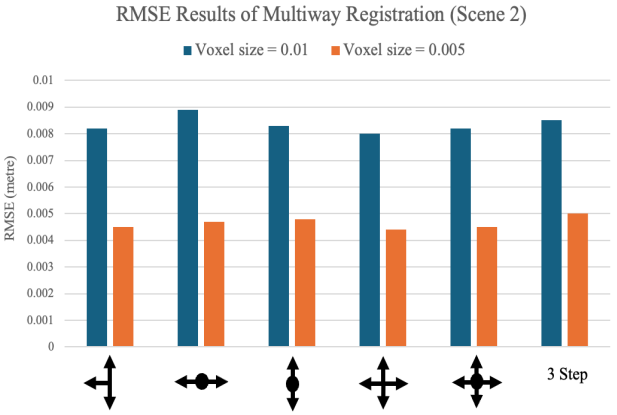

Figure 9. RMSE results of multiway registration for Scene 2

4.2.3 ICP Registration

The RMSE results of ICP registration in Scene 1, as presented in Figure 10, demonstrate registration accuracy based on the different camera movements and voxel sizes used. For a voxel size of 0.01 metres, the highest RMSE value was recorded for the "3 Step" combination, with a value of 0.0714. This suggests that incorporating multiple directional movements introduces more variability, leading to increased registration errors. The "Up, Left, Down" combination achieved competitive results with an RMSE value of 0.0451 for a voxel size of 0.005 metres, which is lower than the single image registration RMSE value of 0.0459. This indicates that incorporating these directional movements enhanced the overall accuracy of the ICP registration in Scene 1.

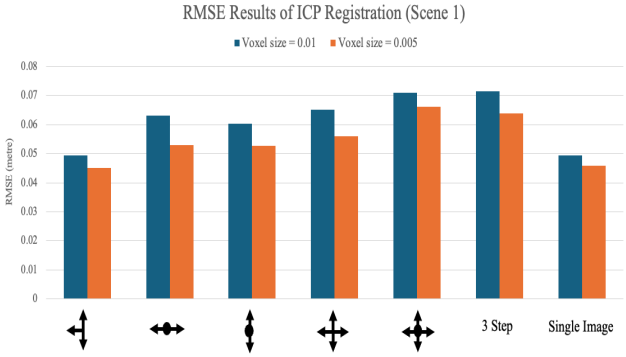

Figure 10. RMSE results of ICP registration for Scene 1

Figure 11 illustrates the RMSE results of ICP registration for Scene 2. For a voxel size of 0.01 metres, the highest RMSE value was recorded for the "Up, Right, Down" combination at 0.0397. For the voxel size of 0.005 metres, the RMSE result for the "Original image, Left, Right" combination was competitive, with an RMSE value of 0.0345, which is just 0.0003 lower than the RMSE obtained from the single image registration. However, the results for Scene 2 indicate that incorporating up and down movements tends to increase the RMSE values, suggesting that vertical shifts introduce more variability and complexity in ICP registration process.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-453-2025 | © Author(s) 2025. CC BY 4.0 License.

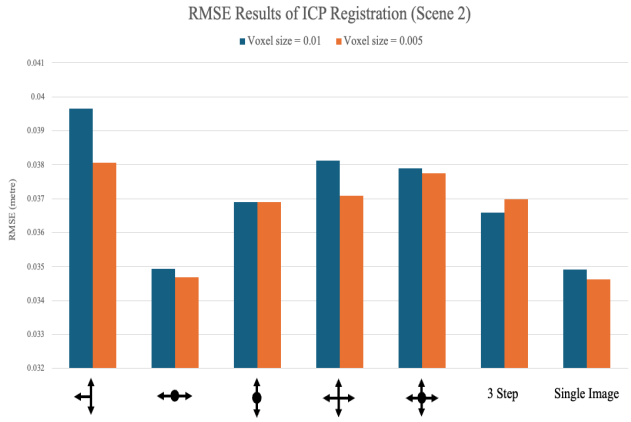

Figure 11. RMSE results of ICP registration for Scene 2

5. Conclusion

This study introduced a novel approach for I2P registration using CMG and MDE, leveraging the MotionCtrl framework and the Depth Anything V2 model. Through the simulation of four main camera movements (Up, Right, Left, and Down) from a single image, this approach generated point clouds with diverse viewpoints. The findings demonstrated that incorporating both vertical and horizontal movements can lead to competitive results, with our approach achieving lower RMSE values than the original image in some scenarios. However, this study has some limitations:

- **Computational complexity:** The multiway registration process is computationally intensive for large-scale environments, especially as the number of point clouds to be registered increases. This intensifies the processing time and resource requirements, making it more challenging to handle complex scenes efficiently.
- **Manual initial alignment:** The approach requires an initial manual alignment, which may increase the time required and be prone to human error, especially for complex scenes.
- **Motion generation for complex edge features and depth:** The success of the generated movements is influenced by the complexity of edge features and depth variations within the scene, potentially impacting the overall registration accuracy.

Based on these limitations, future works can focus on eliminating the need for initial manual alignment by developing fully automated registration techniques, potentially using machine learning or computer vision methods to enhance accuracy and reduce human error. Additionally, integrating semantic context and large language models into the process can be explored to improve CMG in complex scenes.

6. Acknowledgements

This research was supported by the Australian Research Council’s Industrial Transformation Research Programme and DECRA Fellowship [grant numbers: IH210100048, DE220100094]. The authors acknowledge the support of industry partners: Emerson and Rockfield.

References

Bai, C., Fu, R. & Gao, X., 2024. Colmap-pcd: An open-source tool for fine image-to-point cloud registration. 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 1723-1729.

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P. & Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V. & Letts, A., 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Cabon, Y., Murray, N. & Humenberger, M., 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*.

Campbell, D., Liu, L. & Gould, S., 2020. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 244-261.

Chen, Y. & Medioni, G., 1992. Object modelling by registration of multiple range images. *Image and vision computing,* 10**,** 145-155.

Choi, S., Zhou, Q.-Y. & Koltun, V., 2015. Robust reconstruction of indoor scenes. Proceedings of the IEEE conference on computer vision and pattern recognition. 5556-5565.

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D. & Dai, B., 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Gupta, S., Keshari, A. & Das, S., 2022. Rv-gan: Recurrent gan for unconditional video generation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024-2033.

Hartley, R. & Zisserman, A., 2003. *Multiple view geometry in computer vision*, Cambridge university press.

Jeon, Y. & Seo, S.-W., 2022. Efghnet: A versatile image-to-point cloud registration network for extreme outdoor environment. *IEEE Robotics and Automation Letters,* 7**,** 7511-7517.

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. *arXiv preprint arXiv:2309.14660*.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C. & Schindler, K., 2024. Repurposing diffusion-based image generators for monocular depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9492-9502.

Li, J. & Lee, G. H., 2021. DeepI2P: Image-to-point cloud registration via deep classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15960-15969.

Li, J., Li, D., Savarese, S. & Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. International conference on machine learning. PMLR, 19730-19742.

Liu, L., Campbell, D., Li, H., Zhou, D., Song, X. & Yang, R., 2020. Learning 2d-3d correspondences to solve the blind perspective-n-point problem. *arXiv preprint arXiv:2003.06752*.

Lyu, Y., Liang, P. P., Pham, H., Hovy, E., Póczos, B., Salakhutdinov, R. & Morency, L.-P., 2021. StylePTB: A compositional benchmark for fine-grained controllable text style transfer. *arXiv preprint arXiv:2104.05196*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F. & El-Nouby, A., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. & Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence,* 44**,** 1623-1637.

Ren, S., Zeng, Y., Hou, J. & Chen, X., 2022. CorrI2P: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology,* 33**,** 1198-1208.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R. & Susskind, J. M., 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. Proceedings of the IEEE/CVF international conference on computer vision. 10912-10922.

Sheng, Y., Zhang, L., Li, X., Duan, Y., Zhang, Y., Zhang, Y. & Ji, J., 2024. Rendering-Enhanced Automatic Image-to-Point Cloud Registration for Roadside Scenes. *arXiv preprint arXiv:2404.05164*.

Tulyakov, S., Liu, M.-Y., Yang, X. & Kautz, J., 2018. Mocogan: Decomposing motion and content for video generation. Proceedings of the IEEE conference on computer vision and pattern recognition. 1526-1535.

Voleti, V., Yao, C.-H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R. & Jampani, V., 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*.

Wang, B., Chen, C., Cui, Z., Qin, J., Lu, C. X., Yu, Z., Zhao, P., Dong, Z., Zhu, F. & Trigoni, N., 2021. P2-net: Joint description and detection of local features for pixel and point matching. Proceedings of the IEEE/CVF International Conference on Computer Vision. 16004-16013.

Wang, H., Liu, Y., Wang, B., Sun, Y., Dong, Z., Wang, W. & Yang, B., 2023. FreeReg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. *arXiv preprint arXiv:2310.03420*.

Wang, S., Xieshi, M., Zhou, Z., Zhang, X., Liu, X., Tang, Z., Dai, Y., Xu, X. & Lin, P., 2022. Two-channel vae-gan based image-to-video translation. International Conference on Intelligent Computing. Springer, 430-443.

Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D. & Zhou, J., 2024a. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems,* 36.

Wang, Y., Bilinski, P., Bremond, F. & Dantcheva, A., 2020. Imaginator: Conditional spatio-temporal gan for video generation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 1160-1169.

Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P. & Shan, Y., 2024b. Motionctrl: A unified and flexible motion controller for video generation. ACM SIGGRAPH 2024 Conference Papers. 1-11.

Xu, X., Wang, Y., Wang, L., Yu, B. & Jia, J., 2023. Conditional temporal variational autoencoder for action video prediction. *International Journal of Computer Vision,* 131**,** 2699-2722.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J. & Zhao, H., 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10371-10381.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. & Zhao, H., 2024b. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.

Yao, G., Xuan, Y., Chen, Y. & Pan, Y., 2024. Quantity-Aware Coarse-to-Fine Correspondence for Image-to-Point Cloud Registration. *IEEE Sensors Journal*.

Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X. & Shen, C., 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. Proceedings of the IEEE/CVF International Conference on Computer Vision. 9043-9053.

Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D. & Zhou, J., 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhou, Q.-Y., Park, J. & Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*.

Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J. & Sun, T., Lafite: Towards language-free training for text-to-image generation, 2021. *URL https://arxiv.org/abs/2111.13792*.