GraphI2P: Image-to-Point Cloud Registration with Exploring Pattern of Correspondence via Graph Learning

Supplementary Material

1. Experimental Setting

Due to the Image-to-Point Cloud registration performance being heavily affected by the experimental settings, we demonstrate the details of our experimental settings and re-implement details of other methods.

1.1. Datasets Pre-processing Details

There are several widely used public datasets such as KITTI Odometry [1], nuScenes [2], and Oxford RobotCar [3] for cross-modality Image-to-Point Cloud registration even though none of them are designed for this task specifically. Given the broader range of applications of LiDAR point cloud, which is also more challenging to conduct Image-to-Point Cloud registration, we follow the recent approaches VP2P [4] and CorrI2P [5] settings to perform our experiments on KITTI Odometry [1] and nuScenes [2] datasets.

KITTI Odometry [1]. We follow the train/test split in DeepI2P [6], utilizing the 0-8 sequences for training, and 9-10 for testing. Besides, similar to DeepI2P [6], we generate the **non-synchronized frame pairs by randomly selecting image and point cloud frame within $\pm 10m$. Specifically, we randomly select the image and point cloud pairs in test sequences and keep those initial relative poses within** $20m$**.** The random 2D rotation within $[-2\pi, 2\pi]$ and translation within $\pm 10m$ in each direction is applied to the LIDAR point cloud during evaluation. In summary, the initial rotation is unrestricted within the horizontal plane, and the initial translation can reach up to 20 meters within the horizontal plane. and translation can The image size was set to 160×512 as[4] and the point cloud size to 40960 during training and testing, respectively.

nuScenes [2]. We follow the setting utilized in CorrI2P [5] to build point clouds set at the size of radius $80m$. Besides, we empirically remove the points whose height is more than $2.5m$. Furthermore, we leverage the official SDK to generate non-synchronized image-point cloud pairs, where the image frame and point cloud frame **initial position are within** $\pm 10m$**.** We follow the official data split of nuScenes to utilize 850 scenes for training and 150 scenes for testing. The registration methods are also tested on the image and real point cloud with a large relative pose. The random 2D rotation and translation are applied to the LIDAR point cloud during evaluation, respectively. During training, we downsample the image resolution to 160×320 as [4] and the point cloud size to 40960 during training and testing, respectively.

1.2. Evaluation Metrics

Evaluation metrics. Following the previous works [6, 7], we evaluate the performance of the Image-to-Point Cloud registration with widely used metrics: Relative Translation Error (RTE), Relative Rotation Error (RRE), Registration Recall (RR). RTE: The mean of relative error in translation vectors as follows:

$$\text{RTE} = \|t_e - t_{gt}\|, \tag{1}$$

where t_{gt} is the ground-truth of the translation vector, and t_e is the estimated translation vector.

RRE: The mean of RRE between the predicted and the ground-truth camera pose as follows:

$$RRE = \sum_{i=1}^{3} |\theta_i|, \tag{2}$$

where $\{\theta_i\}_{i=1}^{3}$ are the Ruler angles of the rotation error matrix ${R_e}^{-1}R_{gt}$, where R_e and R_{gt} are the estimated and ground-truth of the rotation matrix.

RR: The fraction of successful registrations where the RRE is smaller than τ_r and the RTE is smaller than τ_d as follows:

$$\text{RR} = \frac{1}{M}\sum_{i=1}^{M}\left[\text{RRE}_i < \tau_r \wedge \text{RTE}_i < \tau_t\right], \tag{3}$$

where M is the total number of data samples. RRE_i and RTE_t are the relative rotation and translation errors of the ith data, respectively.

To further verify the effectiveness of our proposed correspondence selection method, we utilize the Inlier Ratio (IR) of the correspondences as well. Unlike the single modality registration task, we take the correspondence as the inlier if the distance between the projected point and ground-truth pixel is smaller than a threshold τ_d as follows,

$$IR = \frac{1}{\left|\tilde{C}\right|}\sum_{(u_i,p_i)\in\tilde{C}}\left[\left[\left\|u_i - F(K\bar{T}p_i)\right\| \leq \tau_d\right]\right], \tag{4}$$

where $[\![\cdot]\!]$ is the Iversion bracket and $\tilde{C}$ is the estimated correspondences set, $\bar{T}$ is the transformation matrix (i.e., $\bar{T} = [\bar{R}|\bar{\text{t}}]$), and K is the camera intrinsic.

We follow the VP2P [4] evaluation metrics to report the RTE, RRE, and RR performances **rather than** to follow the CorrI2P [5] settings, which removes the image-point cloud

Table S1. The effect of depth estimation methods on the KITTI datasets. Lower is better for RTE, RRE, and RMSE, and higher is better for RR and threshold accuracy ($\delta < 1.25$). The best results are indicated in bold, and the second bests are underlined.

Method	Depth Estimation Performance			Registration Performance		
	RMSE↓	Sq. Rel.↓	$\delta < 1.25(\%) \uparrow$	$RTE(m) \downarrow$	$RRE(^\circ) \downarrow$	RR(%) ↑
Baseline	2.28	0.18	97.10	0.86 ± 1.90	2.28 ± 2.41	90.07
-BTS[8]	2.58	0.21	95.20	1.35 ± 1.78	2.62 ± 2.44	85.14
-newCRFs[9]	<u>2.23</u>	<u>0.16</u>	<u>97.40</u>	0.82 ± 1.25	2.23 ± 2.39	91.25
-IEbins[10]	**2.01**	**0.14**	**97.90**	0.71 ± 1.21	2.20 ± 2.46	91.48
Full	2.28	0.18	97.10	<u>0.68 ± 0.95</u>	**1.91 ± 2.04**	**95.40**
-BTS[8]	2.58	0.21	95.20	0.92 ± 1.41	2.35 ± 2.97	88.64
-newCRFs[9]	<u>2.23</u>	<u>0.16</u>	<u>97.40</u>	0.70 ± 1.19	<u>2.02 ± 1.73</u>	<u>93.29</u>
-IEbins[10]	**2.01**	**0.14**	**97.90**	**0.57 ± 1.08**	2.08 ± 1.85	93.02

samples with serious errors before averaging. We further report the registration recall as VP2P [4] setting, which is the proportion of fine registrations with RTE $< 2m$, RTE $< 5^\circ$ on the KITTI and nuScenes datasets.

1.3. Re-implementation Details

For a fair comparison, we conduct re-implementation on open-source methods, such as CorrI2P [5], VP2P [4], Freereg [11], RetrI2P [12], CFI2P [7]. For the same-frame settings, we utilize the results reported in the original paper [4, 7, 12] or take the pre-train mode[5, 11] for testing provided by the official. For the non-synchronized random-frame settings, we retrain the model and conduct testing based on the new settings except the VP2P and FreeReg, for these methods only provide the source code, pre-trained models, and pre-processing settings for the KITTI dataset. For the RetrI2P [12], we replace their depth map generation method with the Zoe-depth [13] for a fair comparison. For the indoor Image-to-Point Cloud registration method 2D3D-Matr [14], we utilize the results provided by [7]. For other recent works [15, 16], we do not take into comparison due to these methods' lack of reliable source code.

2. Ablation study

We conduct extensive ablation studies on KITTI dataset with the non-synchronized random-frame settings.

2.1. The Effect of Depth Estimation Methods

To study the impact of the depth estimation methods, we compare four different monocular depth estimation methods for virtual point cloud generation as Table S1 shows. We introduce widely used root mean squared error (RMSE), relative squared error (Sq. Rel.), and threshold accuracy ($\delta < 1.25$) as the metrics to evaluate the performance of the depth estimation methods. We take our proposed method without virtual-spherical representation as the baseline and replace zoe-depth [13] with other depth estimation methods, such as BTS [8], newCRFs [9], and IEbins [10]. Intuitively, a more effective monocular depth estimation method can generate higher-quality virtual point clouds, thereby improving cross-modal registration performance. However, we can observe from Table S1 that when the depth estimation performance reaches a relatively high level, its impact on registration accuracy becomes minimal. Moreover, the virtual-spherical representation can alleviate the impact of the depth estimation effectively.

2.2. The Effect of Log-normal Distribution Approximate

We conduct ablation on the distribution-based adaptive sample module. Compared to modeling the pattern of the LiDAR point cloud as the log-normal distribution, we use a normal distribution to perform parameter estimation for the LiDAR point cloud distribution. We set conducting random sample as the baseline. The experimental results in Table S2 indicate that the log-normal distribution more closely approximates the original distribution of LiDAR points in the depth direction.

Table S2. The effect of Distribution Approximate.

Point	RTE(m)↓	RRE(°)↓	RR(%).↑
Baseline	1.33 ± 1.52	2.61 ± 1.94	84.81
+ Normal	1.30 ± 1.36	2.56 ± 1.87	85.03
+ Log-Normal	**0.68 ± 0.95**	**1.91 ± 2.04**	**95.40**

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 1

[3] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 1

[4] Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[5] Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1198–1208, 2022. 1, 2

[6] Jiaxin Li and Gim Hee Lee. DeepI2P: Image-to-point cloud registration via deep classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15960–15969, 2021. 1

[7] Gongxin Yao, Yixin Xuan, Yiwei Chen, and Yu Pan. Quantity-aware coarse-to-fine correspondence for image-to-point cloud registration. *IEEE Sensors Journal*, 2024. 1, 2

[8] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[9] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022. 2

[10] Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. Iebins: Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[11] Haiping Wang, Yuan Liu, Wang Bing, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *International Conference on Learning Representations*, 2024. 2

[12] Lin Bie, Siqi Li, and Kai Cheng. Image-to-point registration via cross-modality correspondence retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 266–274, 2024. 2

[13] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[14] Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14128–14138, 2023. 2

[15] Gongxin Yao, Xinyang Li, Yixin Xuan, and Yu Pan. Mafreei2p: A matching-free image-to-point cloud registration paradigm with active camera pose retrieval. In *2024 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2024. 2

[16] Shuhao Kang, Youqi Liao, Jianping Li, Fuxun Liang, Yuhao Li, Xianghong Zou, Fangning Li, Xieyuanli Chen, Zhen Dong, and Bisheng Yang. Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics and Automation Letters*, 2024. 2