Supplement Materials of MinCD-PnP: Learning 2D-3D Correspondences with Approximate Blind PnP

Pei An[1]*, Jiaqi Yang[2]*, Muyao Peng[1], You Yang[1]†, Qiong Liu[1], Xiaolin Wu[3], Liangliang Nan[4]
[1]Huazhong University of Science and Technology, China
[2]Northwestern Polytechnical University, China
[3]Southwest Jiaotong University, China [4]Delft University of Technology, Netherlands

1. Implementation details of MinCD-Net

More details of MinCD-Net are discussed here. Its inputs include an RGB image with surface normals and an RGB point cloud with surface normals. Image surface normals are predicted using the pre-trained model DSINE [2]. The extractors are ResNet [5] and KPConv [13], where the extractor networks are similar to those in MATR [8].

The threshold s_{th} in Eq. (14) is set to $e^{-0.4}$. Point transformer is the single layer of work [15]. Its key, query, and value inputs are the 128 dimensional features which are transformed from pixels and points features. To estimate the camera pose, we use two-layer MLPs with dimensions $[256, 128]$ and $[128, 6]$ to predict a 6×1 vector representing the $\mathbf{se}(3)$ of $\mathbf{T}$, and $\mathbf{T}$ is computed via the mapping from $\mathbf{se}(3)$ to $\mathbf{SE}(3)$. We utilize Shi-Tomasi keypoint detection provided by OpenCV API `Good Features to Track` to extract $\mathbf{K}_{\text{I}}$ that are uniformly distributed in the image. We train MinCD-Net on a single NVIDIA RTX 3080 GPU for 40 epochs. To evaluate the proposed method in the practical applications, we prepare a self-collected dataset. It is captured by an Intel RealSense depth camera. Examples of scenes are provided in Fig. 1.

2. Additional comparisons

We evaluate the registration performance of current I2P registration methods on the outdoor KITTI benchmark [4]. DeepI2P [7], CorrI2P [10], VP2P-Match [16], CoFiI2P [6], CMR-Agent [14], and OL-Reg [1] are used for comparison. We utilize the pretrained CorrI2P to preprocess the LiDAR point clouds to gather the overlapped LiDAR point clouds, as the inputs of MATR+MinCD-Net. Results are shown in Table 1. The average relative translational error (RTE) and average relative rotation error (RRE) are used as metrics. These results indicate that MinCD-Net achieves state-of-the-art performance on the KITTI benchmark.

Table 1. Comparisons on the KITTI dataset. † indicates the usage of a pretrained model to segment the overlapped point cloud, which falls into the field of view of the camera.

Methods	Venue	RTE/m	RRE/deg
DeepI2P	CVPR 2021	1.460	4.270
CorrI2P	IEEE T-CSVT 2022	0.740	2.070
VP2P-match	NeurIPS 2023	0.750	3.290
CoFiI2P	IEEE RAL 2024	0.290	1.140
CMR-Agent	IROS 2024	0.195	0.589
MATR+MinCD-Net†		**0.091**	**0.228**

Table 2. Ablation study of the proposed method with different choices of 2D keypoint detectors.

Schemes	Shi-Tomasi (used)	FAST	SIFT	SuperPoint	Uniformly sampled
IR	0.567	0.552	0.572	0.560	0.542
RR	0.646	0.631	0.638	0.649	0.625

Table 3. Computational efficiency analysis of the current methods. Diff. PnP, BPnPNet, and MinCD-Net are only used to supervise the backbone networks (not used in the inference stage), so the runtime and GPU memory in the training stage are recorded.

Methods	Runtime/ms	Param/M	GPU memory/MB	RR
Baseline	127	28.2	7532	51.0%
+Diff. PnP	152 (+25)	28.2 (+0.0)	7852 (+320)	49.1%
+BPnPNet	141 (+14)	30.8 (+2.6)	8242 (+710)	57.8%
+MinCD-Net	148 (+21)	31.4 (+3.2)	8353 (+821)	**64.7%**

Additional qualitative results are shown in Fig. 2. It is found that the proposed MinCD-Net achieves both robust and accurate performance compared to existing differentiable PnP based methods in the cross-scene setting.

3. Additional ablation studies

We investigate the dependency of MinCD-Net on 2D keypoint detectors, like FAST [11], SIFT [9], Superpoint [12], and even the uniformly sampled scheme. Results in Table 2 indicate that MinCD-Net achieves nearly the same results as other standard detectors, even with uniform sampling. This indicates that while MinCD-Net requires 2D key-

*Equal contribution
†Corresponding author: yangyou@hust.edu.cn

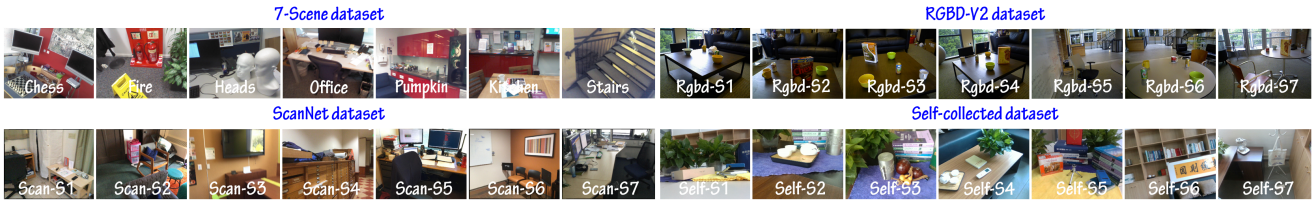

Figure 1. Example scenes from the 7-Scenes, RGBD-V2, ScanNet, and self-collected datasets (referred to as *Rgbd*, *Scan*, and *Self*).

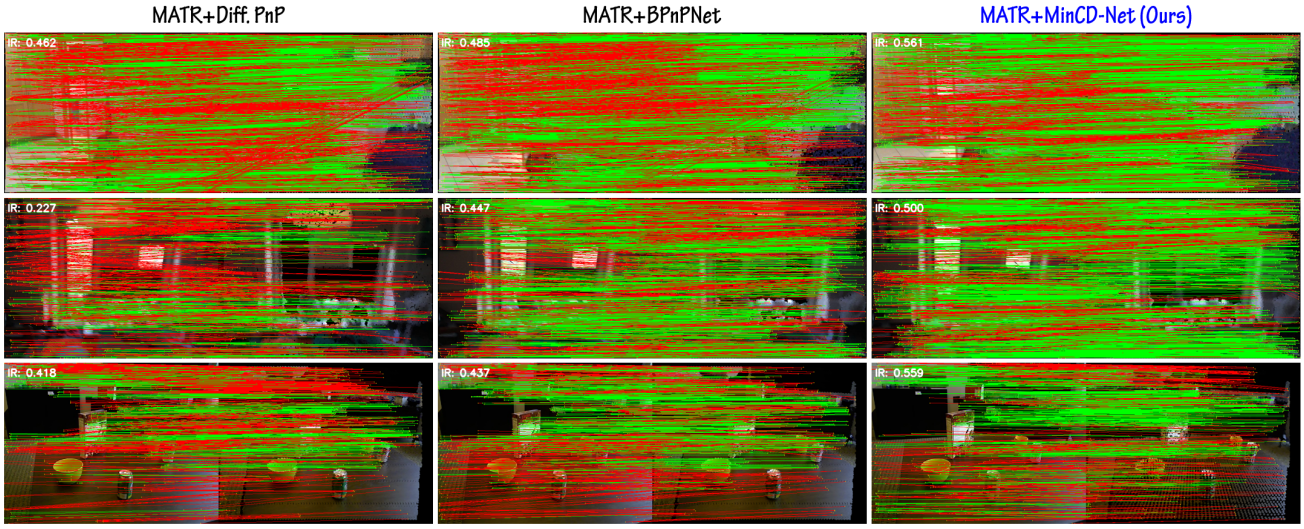

Figure 2. Visualization of different methods. MinCD-Net achieves a higher correspondence accuracy than other methods.

points, it does not depend on a specific detection method. Besides, the computational analysis of the current methods is provided in Table 3. It indicates that MinCD-Net is a lightweight network with comparable runtime and GPU memory. Overall, the above results show the effectiveness of MinCD-Net.

4. Limitations and future work

In the challenging scenarios (e.g., the self-collected dataset), the performance gain of MinCD-Net is limited (as seen in Table 2 in the format conference manuscript). The precision of learned 3D keypoints is not high (as seen in Table 6 in the format conference manuscript). To address these limitations, we plan to integrate a learnable correspondences pruning module [3] to improve the efficiency of solving MinCD-PnP.

References

[1] Pei An, Xuzhong Hu, Junfeng Ding, Jun Zhang, Jie Ma, You Yang, and Qiong Liu. Ol-reg: Registration of image and sparse lidar point cloud with object-level dense correspondences. *IEEE Trans. Circuits Syst. Video Technol.*, 34(8): 7523–7536, 2024. 1

[2] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proc. CVPR*, pages 9535–9545, 2024. 1

[3] Yuxin Cheng, Zhiqiang Huang, Siwen Quan, Xinyue Cao, Shikun Zhang, and Jiaqi Yang. Sampling locally, hypothesis globally: accurate 3d point cloud registration with a ransac variant. *Visual Intelligence*, 20:1–15, 2023. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. CVPR*, pages 3354–3361, 2012. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 1

[6] Shuhao Kang, Youqi Liao, Jianping Li, and et al. Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics Autom. Lett.*, 9(11):10264–10271, 2024. 1

[7] Jiaxin Li and Gim Hee Lee. DeepI2P: Image-to-point cloud registration via deep classification. In *Proc. CVPR*, pages 15960–15969, 2021. 1

[8] Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2D3D-MATR: 2D-3D match-

ing transformer for detection-free registration between images and point clouds. In *Proc. ICCV*, pages 1–10, 2023. 1

[9] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. 1

[10] Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. CorrI2P: Deep image-to-point cloud registration via dense correspondence. *IEEE Trans. Circuits Syst. Video Technol.*, 33(3):1198–1208, 2023. 1

[11] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Proc. ECCV*, pages 430–443, 2006. 1

[12] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. CVPR*, pages 4937–4946, 2020. 1

[13] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proc. ICCV*, pages 6410–6419, 2019. 1

[14] Gongxin Yao, Yixin Xuan, Xinyang Li, and Yu Pan. Cmr-agent: Learning a cross-modal agent for iterative image-to-point cloud registration. In *Proc. IROS*, pages 13458–13465, 2024. 1

[15] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proc. ICCV*, pages 16239–16248, 2021. 1

[16] 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. In *Proc. NeurIPS*, pages 1–10, 2023. 1