Implicit Correspondence Learning for Image-to-Point Cloud Registration

Supplementary Material

1. More implementation details

Here we offer more implementation details about our proposed method. For feature extraction, we adopt backbones commonly used in existing approaches [2]. Specially, we adopt a 4-stage ResNet [1] with a FPN [3] as the image backbone, where the output channel dimension is 128. The output feature map has the same resolution as the input image, which is 160×512 for the KITTI dataset and 160×320 for the nuScenes dataset. For the 3D backbone, we use a 4-stage KPFCNN [5] where the output channel dimension is 128. The point clouds are voxelized with an initial voxel size of 15cm for the KITTI dataset and 30cm for the nuScenes dataset. The batch size is set as 2. All experiments are conducted on a single RTX 3090 GPU. We implement our code using PyTorch 1.13.1 and CUDA 11.7.

2. Inverse-projection of the 2D keypoints

In the pose regression module, the 2D keypoints $\mathbf{K}_I$ are first inverse-projected onto the camera intrinsic-independent space, *i.e.*, the normalized plane of the camera pinhole model. The 2D keypoints are inverse-projected as:

$$[\overline{u_i}, \overline{v_i}, 1]^\top = \mathbf{K}^{-1}[u_i, v_i, 1]^\top, \tag{1}$$

where $\mathbf{K}$ is the intrinsic matrix of the camera, $[u_i, v_i]^\top$ is the pixel coordinate of the i-th 2D keypoint.

3. Evaluation metrics

Relative Translational Error (RTE). RTE is a metric used to quantify the accuracy of estimated translations in 3D space. It computes the discrepancy between the estimated translation vector and ground-truth translation vector. RTE is defined as the Euclidean distance between the two translation vectors:

$$\mathbf{RTE} = \|\mathbf{t}_E - \mathbf{t}_{gt}\|_2, \tag{2}$$

where $\mathbf{t}_E$ is the estimated translation vector, $\mathbf{t}_{gt}$ is the ground-truth translation vector.

Relative Rotation Error (RRE). RRE is a metric used to quantify the accuracy of estimated rotations in 3D space. It computes the angular difference between the estimated rotation and ground-truth rotation. The error is calculated using the Euler angles representation, by first determining the rotation that transforms the estimated rotation into the ground-truth rotation. RRE is then obtained as the total angular displacement between the two rotations, typically expressed in degrees:

$$\mathbf{RRE} = \sum_{i=1}^{3} |\gamma(i)|\,, \tag{3}$$

where γ is the Euler angles of the matrix $\mathbf{R}_E^{-1}\mathbf{R}_{gt}$, $\mathbf{R}_E$ is the estimated rotation matrix, $\mathbf{R}_{gt}$ is the ground-truth rotation matrix.

4. Quantitative comparisons of correspondence learning

To demonstrate the effectivity of our proposed implicit correspondence learning module, we show the quantitative results of the pixel-point matching accuracy. We mainly compare our method with VP2P-match [6] on the KITTI dataset using 2 metrics: Feature Matching Recall and Projection Error.

Feature Matching Recall (FMR) is the fraction of image-point-cloud pairs whose ratio of correctly estimated correspondences is over a threshold $\tau_1 = 0.95$. A correspondence is regarded as correctly matched if its ground-truth 2D distance is smaller than τ_2 pixels:

$$\mathbf{FMR} = \frac{1}{M}\sum_{i=1}^{M}[\mathbf{IR}_i > \tau_1], \tag{4}$$

$$\mathbf{IR}_i = \frac{1}{|C_i|}\sum_{(x_k, y_k)\in C_i}[\|\mathcal{K}(\mathbf{R}_{gt}\cdot x_k + \mathbf{t}_{gt}) - y_k\|_2 < \tau_2], \tag{5}$$

where $[\cdot]$ denotes the Iverson bracket, M denotes the number of all image-point-cloud pairs, x_k denotes a 3D point, y_k denotes a 2D pixel, C_i denotes all the corresponding pairs of the i-th image-point-cloud pair, $\mathbf{R}_{gt}$ and $\mathbf{t}_{gt}$ denote the ground-truth camera pose, $\mathcal{K}(\cdot)$ projects a 3D point to a 2D pixel according to the camera intrinsic matrix $\mathbf{K}$.

Projection Error (PE) is the ground-truth 2D distance of a correspondence:

$$\mathbf{PE} = \|\mathcal{K}(\mathbf{R}_{gt}\cdot x_k + \mathbf{t}_{gt}) - y_k\|_2. \tag{6}$$

As shown in Table 1, we report the quantitative results under different settings. Our method outperforms VP2P-match [6] in all metrics by a large margin, which demonstrates that our proposed implicit correspondence learning can achieve superior performance than explicit matching strategies in previous methods [4, 6].

Table 1. Quantitative comparisons of correspondence learning on the KITTI dataset.

Methods	FMR(%)				PE (pixels)
	$\tau_2 = 5$	$\tau_2 = 10$	$\tau_2 = 15$	$\tau_2 = 20$	
VP2P-match [6]	19.52	59.76	77.53	85.39	7.43±31.31
Ours	**33.58**	**68.46**	**82.04**	**88.97**	**6.16±7.54**

5. Keypoint heatmaps visualization

To demonstrate the effectivity of our proposed implicit correspondence learning module, we visualize the 2D keypoint heatmap $\mathbf{H}_I$ and 3D keypoint heatmap $\mathbf{H}_P$ in Figure 1. Specially, we draw the heatmaps generated from the same query across different scenes as well as heatmaps from different queries on the same scene. For intuitive visualization, we project the point cloud into image space through ground-truth transformations. In addition, we explicitly mark the corresponding 2D keypoints $\mathbf{K}_I$ and 3D keypoints $\mathbf{K}_P$ on the visualizations, highlighting their spatial alignment and the learned correspondences across 2D and 3D domains. As shown in the Figure 1, the heatmaps from different queries can focus on different regions of the input scene to comprehensively depict the whole scene. Also, the heatmaps from the same query tends to focus on regions with similar structures and similar spatial locations within the images across different scenes, which demonstrates the generalization ability of our generated detectors across different scenes. Marked keypoints indicate that the 2D keypoint and 3D keypoint generated from the same query correspond to the same object, which demonstrates that our proposed implicit correspondence learning module can obtain accurate 2D-3D correspondences.

6. More qualitative comparisons

Here, we provide more qualitative comparisons between our proposed method and VP2P-match [6] in Figure 2, highlighting the differences in registration accuracy and visual alignment across various challenging scenarios. The figure demonstrates our method consistently outperforms VP2P-match [6] in terms of accuracy and robustness, particularly in scenes with complex geometries.

7. Discussion of generalizability

Though our method is proposed to address the challenge of cross-modality matching while trying to figure out the image-to-point cloud registration problem, we believe our method also holds potential in the field of image matching. In general scenarios, our method may not perform as well as existing image matching methods, because the feature differences between images are relatively small and making direct matching is practical. However, while dealing with image pairs that have significant domain differences, our method may perform better. We may conduct relevant experiments to validate our idea in our future work.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] 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. 1

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[4] 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

[5] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 1

[6] 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 *Advances in Neural Information Processing Systems*, pages 51166–51177. Curran Associates, Inc., 2023. 1, 2, 3

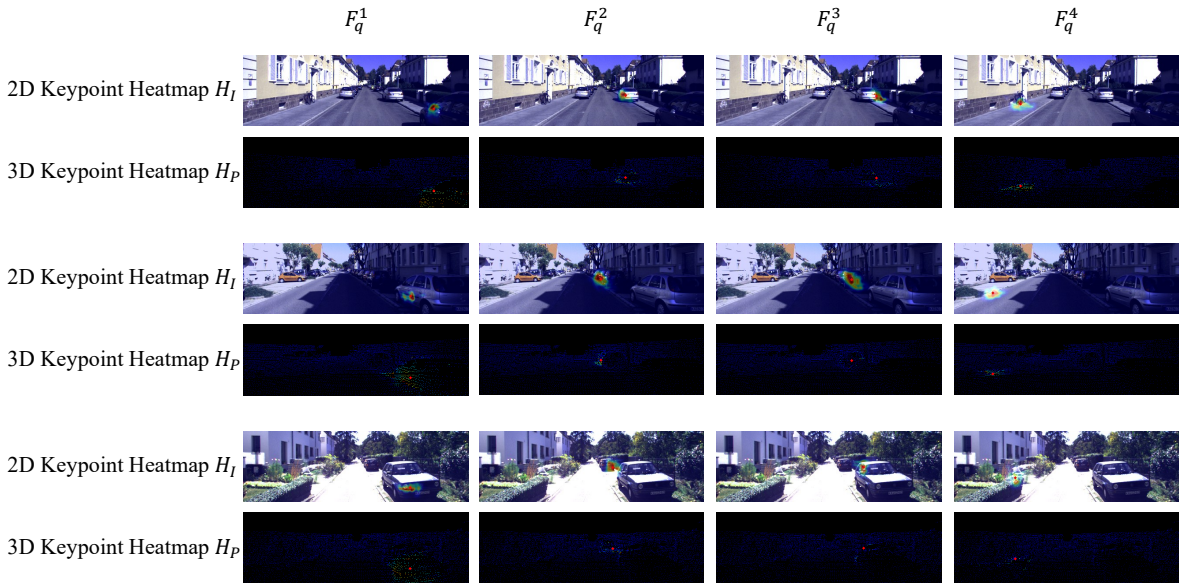

Figure 1. Visual illustration of 2D/3D keypoint heatmaps and corresponding 2D/3D keypoints. Each row represents the heatmaps generated from different queries on a same scene. Each column represents the heatmaps generated from a same query across different input scenes. Red/blue indicates a large/small weight. The keypoints are marked as red diamond-shaped symbols.

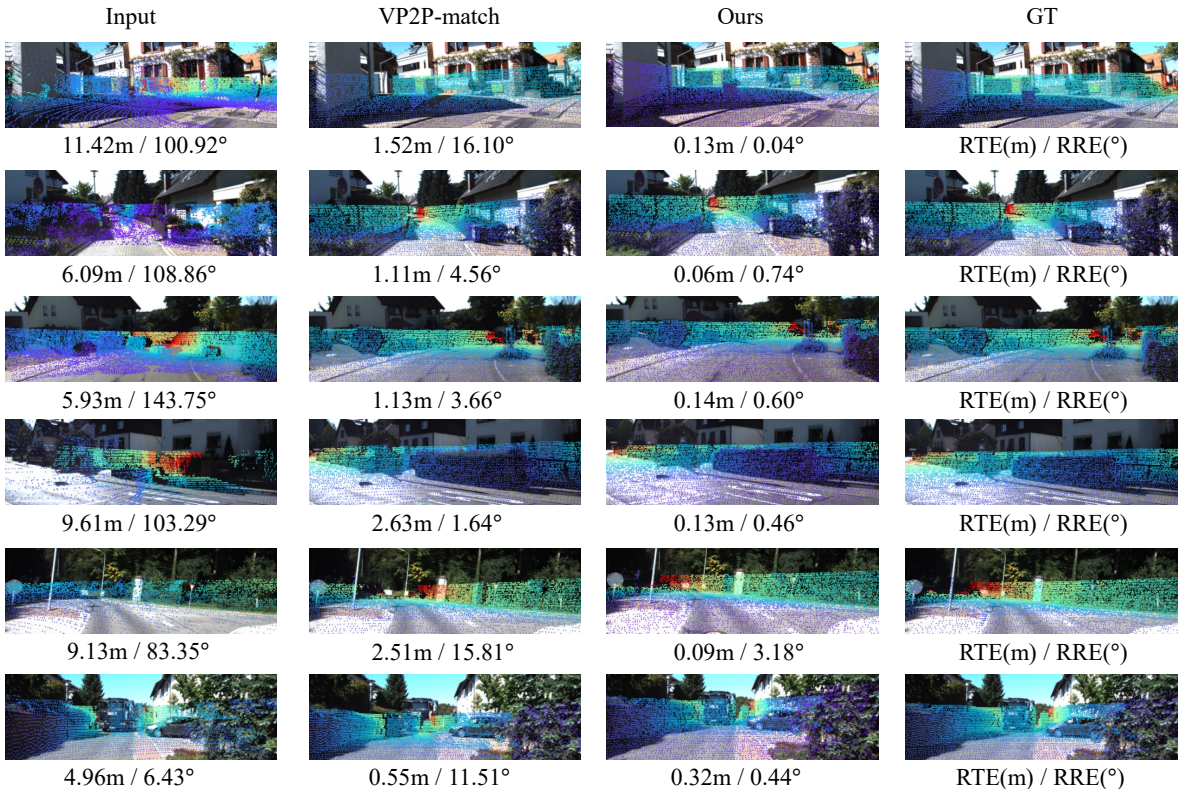

Figure 2. More qualitative comparisons between our proposed method and VP2P-match [6].