Supplementary Material:

SIRA-PCR: Sim-to-Real Adaptation for 3D Point Cloud Registration

Suyi Chen¹,⁴*  Hao Xu²*  Ru Li³  Guanghui Liu¹†  Chi-Wing Fu²  Shuaicheng Liu¹,⁴†

¹University of Electronic Science and Technology of China  ²The Chinese University of Hong Kong  ³Harbin Institute of Technology  ⁴Megvii Technology

*Equal contribution.  †Corresponding authors.
This supplementary material document contains the following two parts.

**Part 1** presents details of the evaluation metrics, additional evaluation results, and detailed network architecture of SIRA.

**Part 2** presents more visualization and qualitative results.
Part 1.1: Evaluation Metrics in Experiments

Following the existing works, we present the definitions of the evaluation metrics we employed.

(i) **Registration Recall (RR)** is the fraction of successfully registered point cloud pairs. A point cloud pair is said to be successfully registered when its transformation error is lower than threshold \( \tau \).

In addition, the transformation error is defined as the root mean square error of the ground-truth correspondences \( C^* \), to which the estimated transformation \( T_{est}(\cdot) \) has applied:

\[
RMSE = \sqrt{\frac{1}{|C^*|} \sum_{(p^*_x, q^*_y) \in C^*} ||T_{est}(p^*_x) - q^*_y||^2_2},
\]

\[
RR = \frac{1}{M} \sum_{i=1}^{M} [RMSE_i < \tau_1],
\]

where \( p_x \) and \( q_y \) denote the \( x \)-th point in source \( P \) and \( y \)-th point in target \( Q \), respectively; \([ \cdot ]\) is the Iversion bracket; and \( M \) is the number of all point cloud pairs.

(ii) **Inlier Ratio (IR)** is the fraction of inlier correspondences among all hypothesized correspondences \( C \). A correspondence is regarded as an inlier if the distance between the two points is lower than a certain threshold \( \tau_2 \) under the ground-truth transformation \( T_{gt}(\cdot) \):

\[
IR = \frac{1}{|C|} \sum_{(p_x, q_y) \in C} [||T_{gt}(p_x) - q_y||_2 < \tau_2].
\]

(iii) **Feature Matching Recall (FMR)** is the fraction of point cloud pairs whose IR > threshold \( \tau_3 \):

\[
FMR = \frac{1}{M} \sum_{i=1}^{M} [IR_i > \tau_3].
\]

(iv) **Relative Rotation Error (RRE)** is the geodesic distance in degrees between the estimated and ground-truth rotation matrices \( R_{est} \) and \( R_{gt} \):

\[
RRE = \arccos \left( \frac{\text{trace}(R_{est}^T \cdot R_{gt} - 1)}{2} \right).
\]

(v) **Relative Translation Error (RTE)** is the Euclidean distance between estimated and ground-truth translation vectors \( t_{est} \) and \( t_{gt} \):

\[
RTE = ||t_{est} - t_{gt}||_2.
\]

(vi) **Transformation Recall (TR)** is the fraction of successfully registered point cloud pairs, similar to RR but using a different definition of successful registration:

\[
TR = \frac{1}{M} \sum_{i=1}^{M} [RRE_i < \tau_4 \text{ and } RTE_i < \tau_5].
\]

TR is also called RR in [1, 11], while we follow [13] to use TR to avoid ambiguity.

Following [2, 7, 11, 15], we set \( \tau_1 = 0.2m \), \( \tau_2 = 0.1m \), \( \tau_3 = 0.05 \), \( \tau_4 = 15^\circ \), and \( \tau_5 = 0.3m \) for evaluation on the 3DMatch and 3DLoMatch benchmarks. Following [13], we set \( \tau_1 = 0.5m \), \( \tau_2 = 0.2m \), and \( \tau_3 = 0.05 \) for evaluation on the ETH benchmark.
Part 1.2: More Experimental Results

In this section, we present additional experiments that evaluate our method. Note that we present these additional experiments because different existing methods use different evaluation metrics/approaches.

Following [7, 11], we evaluate the scene-wise registration on the 3DMatch and 3DLoMatch datasets. Tab. 1 reports the results. Since extremely large errors generated from failure cases can easily dominate the results, median RRE and RTE for each scene are reported only on the successfully-registered pairs. As Tab. 1 shows, although methods with higher RR tend to generate higher RRE and RTE since more challenging cases are taken into consideration, we can see that our method always has better improvements in RRE and RTE regardless of RR (no matter higher/lower than the competitors); see again Tab. 3. The robustness shown in this experiment is mainly attributed to the method always has better improvements in RRE and RTE regardless of RR (no matter higher/lower than the competitors) since it outperforms Lepard by 2.5/3.3 pp on 3DMatch/3DLoMatch with lower RRE and RTE, showing the best performance (see the “mean” columns) for all scenes and datasets.

For a fair comparison with Lepard [9], we also report the results using the metrics following them. Here, RR is averaged on all scan pairs, different from the one averaged over scenes in Tab. 1. As Tab. 2 shows, our method achieves the best performance on all metrics for both datasets. Specifically, it outperforms Lepard by 2.5/3.3 pp on 3DMatch/3DLoMatch with lower RRE and RTE, showing the robustness and effectiveness of our method.

Following OIF-PCR [14], we report the mean median RRE and RTE obtained by RANSAC for the successfully-registered scan pairs. As Tab. 3 shows, although methods with higher RR tend to generate higher RRE and RTE since more challenging cases are taken into consideration, we can see that our method always has better improvements in RRE and RTE regardless of RR (no matter higher/lower than the competitors); see again Tab. 3. The robustness shown in this experiment is mainly attributed to the
Absolutely accurate annotations generated by our approach.

<table>
<thead>
<tr>
<th>Model</th>
<th>3DMatch</th>
<th>3DLoMatch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RRE ↓</td>
<td>RR ↑</td>
</tr>
<tr>
<td></td>
<td>RTE ↓</td>
<td>RR ↑</td>
</tr>
<tr>
<td>PREDATOR [7]</td>
<td>2.72</td>
<td>91.8</td>
</tr>
<tr>
<td>Lepard [9]</td>
<td>2.48</td>
<td>93.5</td>
</tr>
<tr>
<td>PREDATOR [7] + ICP</td>
<td>2.06</td>
<td>92.3</td>
</tr>
<tr>
<td>Lepard [9] + ICP</td>
<td>1.96</td>
<td>93.9</td>
</tr>
<tr>
<td>GeoTrans [11]</td>
<td>1.84</td>
<td>94.1</td>
</tr>
</tbody>
</table>

Table 2. Registration results on 3DMatch and 3DLoMatch following the metrics used in [9]. †: the model is only trained on the synthetic dataset. ‡: the model is fine-tuned on 3DMatch.

In [6], an unsupervised domain adaptation approach, named UDGE, has been proposed to apply cubic cropping on 3DMatch to generate pairs. However, as shown in Tab. 5, UDGE shows limited gains compared with the baseline when it is applied to domain adaptation from our synthetic indoor dataset FlyingShapes to the real-world indoor dataset 3DMatch. Since generated point cloud pairs have the same coordinates in overlap regions, it may cause over-fitting of the feature extractor during training, leading to performance degradation. The results also show the necessity of the sim-to-real adaptation (SIRA) in our pipeline.

In Tab. 6, we present supplementary ablation results on FMR and IR metrics. Our SIRA improves performance on all metrics consistently, showing its effectiveness.
<table>
<thead>
<tr>
<th>Model</th>
<th>3DMatch</th>
<th>3DLoMatch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RRE ↓</td>
<td>RFE ↓</td>
</tr>
<tr>
<td>(a) baseline</td>
<td>1.672</td>
<td>0.051</td>
</tr>
<tr>
<td>(b) UEDGE [6]</td>
<td>1.710</td>
<td>0.050</td>
</tr>
<tr>
<td>(c) SIRA†</td>
<td>1.609</td>
<td>0.052</td>
</tr>
</tbody>
</table>

Table 5. Comparison with another UDA method in [6].

<table>
<thead>
<tr>
<th>Model</th>
<th>3DMatch</th>
<th>3DLoMatch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FMR ↑</td>
<td>IR ↑</td>
</tr>
<tr>
<td>FlyingShapes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) Structured3D</td>
<td>94.1</td>
<td>49.8</td>
</tr>
<tr>
<td>(b) 3DFront</td>
<td>87.9</td>
<td>42.7</td>
</tr>
<tr>
<td>(c) + ShapeNet</td>
<td>93.8</td>
<td>57.0</td>
</tr>
<tr>
<td>(d) + Delete planes</td>
<td>95.6</td>
<td>59.7</td>
</tr>
<tr>
<td>(e) + Structured3D</td>
<td>95.5</td>
<td>59.5</td>
</tr>
</tbody>
</table>

| SIRA |       |       |       |
| (a) baseline | 95.5 | 59.5 | 87.5 | 77.8 | 30.6 | 63.8 |
| (b) uniform noise | 96.3 | 62.5 | 88.6 | 77.2 | 30.7 | 62.7 |
| (c) Gaussian noise | 97.1 | 63.9 | 89.3 | 78.6 | 32.6 | 63.8 |
| (d) TSDF | 95.3 | 60.1 | 87.9 | 78.3 | 31.0 | 63.9 |
| (e) w/o ARM | 97.3 | 65.4 | 90.8 | 81.7 | 36.2 | 65.3 |
| (f) w/o multi-scale | 97.5 | 65.3 | 91.2 | 80.6 | 35.8 | 66.6 |
| (g) SIRA† | 97.7 | 65.5 | 91.5 | 81.7 | 35.8 | 67.4 |

Table 6. Supplementary results of ablation studies.

**Part 1.3: Network Architecture Details**

Fig. 1 depicts the detailed network architecture of SIRA. The generator of SIRA is an encoder-decoder structure based on the KPConv backbone [12]. Group normalization and LeakyReLU are applied after each KPConv layer. To better capture the point pattern distributions, the first KPConv layer is applied to the point clouds with 30,000 points, which are randomly sampled rather than voxel downsampling from the original dense point clouds. The $l$-th strided convolution is applied to a voxel-downsampled point cloud with voxel size $0.025 \cdot 2^l$ m. The upsampling module in the generator is performed by searching the corresponding feature of the closest point from the previous layer. ARMs are applied to each upsampling step to provide adaptive point positions. The discriminator is a multi-scale architecture. Encoders based on the same simplified PointNet structure [10], i.e., a shared MLP followed by max-pooling, first extracts the features of multi-scale local patches with 5/10/20 points. The features are then concatenated together, being fed into the subsequent 5-layer shared MLP for classification. For more details, please see our code.

**Part 1.4: Details in evaluation on ETH**

There are many differences, such as the density and the scale, between the point clouds from indoor 3DMatch and outdoor ETH. Thus, it is improper for models totally trained on indoor datasets to be directly evaluated on outdoor benchmarks. For this reason, following [13], we scale the point cloud from ETH to 1/10 of its original size before inference. Note that the point cloud will be rescaled to its original size when computing metrics for a fair comparison.

**Part 2: More Visualization and Qualitative Results**

For better observation, we provide some visualization results in the video in our supplementary material. Layouts of 10 selected indoor scenes in FlyingShapes are displayed, including depth and segmentation images from different viewpoints. We also provide more qualitative results on 3DMatch and 3DLoMatch benchmarks in our video. The results from GeoTrans [11] and RegTR [15] are shown for comparison. More qualitative results on the ETH benchmark are presented in the video as well.
Figure 1. Architecture details of SIRA.

References


