SemReg: Semantics Constrained Point Cloud Registration
Supplementary Material

Sheldon Fung[1], Xuequan Lu*[1], Dasith de Silva Edirimuni[2], Wei Pan[3], Xiao Liu[2], and Hongdong Li[4]

[1] La Trobe University, Australia
[2] Deakin University, Australia
[3] OPTMV, China
[4] Australian National University, Australia

1 Network Configuration

1.1 Local Feature Encoder

Our local feature encoder network comprises a series of residual convolution layers and multiple downsampling layers. A detailed illustration of the network is shown in Figure 1.

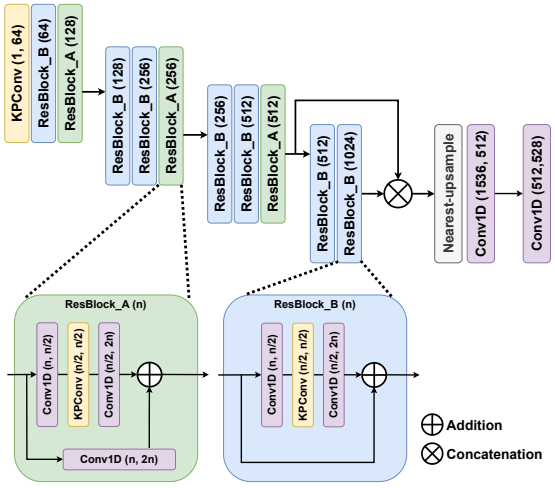

Fig. 1: Illustration of our local feature encoder network. $(\cdot, \cdot)$ indicates the number of input and output channels of the corresponding block. $\oplus$ and $\otimes$ represent addition and concatenation operations, respectively.

* Corresponding author.

1.2 Semantics Guided Feature Interaction

In the Semantics Guided Feature Interaction (SGFI) module, the semantic similarity threshold η is set to 0.1. In the self/cross-attention layers, we set the feature dimension to 528 and the number of heads to 4.

1.3 Semantics Constrained Matching

The Semantics Constrained Matching (SCM) module takes the point-wise features from the SGFI module as input. We thus set feature dimension d to be the same as the SGFI module (i.e., 528). And we set the threshold τ_m to 0.3 in the testing stage.

1.4 Training Details.

Our method is implemented in PyTorch and we use an SGD optimizer with an initial learning rate of 5×10^{-2} and a decay rate of 0.95 for every epoch. The weight decay is set to 1×10^{-6}. Training is conducted on 6 NVIDIA A100 GPUs. We train the network for 40 epochs with a batch size of 1, requiring approximately 24 hours.

2 Evaluation Metrics

We follow recent state-of-the-art approaches [3, 5] and report Inlier Ratio (IR), Feature Matching Recall (FMR), Registration Recall (RR) on the 3DMatch [7] evaluation dataset. For fine alignment results, we follow [4] and report Relative Rotation Error (RRE) and Relative Translation Error (RTE) on top of RR.

The Inlier Ratio (IR) is the fraction of inlier matches among all predicted matches. A match is considered to be an inlier match if the distance between the two points is smaller than $\tau_1 = 0.1m$ under the ground-truth transformation T_g:

$$\mathrm{IR} = \frac{1}{|C_p|} \sum_{(x_i, y_i) \in C_p} [\![\|T_g(x_i) - y_i\|_2 < \tau_1]\!], \tag{1}$$

where C_p is a set of predicted matches and $[\![\cdot]\!]$ is the Iverson bracket.

Feature Matching Recall (FMR) is the fraction of point cloud pairs whose IR is above $\tau_1 = 0.05$. It indicates the potential for successful registration:

$$\mathrm{FMR} = \frac{1}{N} \sum_{i=1}^{N} [\![\mathrm{IR}_i > \tau_2]\!], \tag{2}$$

where N is the number of all point cloud pairs in the dataset.

Registration Recall (RR) is the fraction of the point cloud pair that is correctly registered among all samples in the dataset. A pair of point clouds is successfully registered if the transformation error e_i is smaller than $0.2m$:

$$\mathrm{RR} = \frac{1}{N}\sum_{i=1}^{N}(e_i < 0.2m), \tag{3}$$

where N is the number of the samples in the dataset. The transformation error e_i is defined as the Root Mean Square Error (RMSE) of the ground truth correspondence $\mathbb{C}^*$ after applying the estimated transformation:

$$e_i = \sqrt{\frac{1}{|\mathbb{C}^*|}\sum_{(\mathbb{C}_{xi},\mathbb{C}_{yi})\in\mathbb{C}^*}\|T_g(\mathbb{C}_{xi}) - \mathbb{C}_{yi}\|_2^2}, \tag{4}$$

where T_g is the ground-truth transformation that aligns the point cloud pair.

Relative Translation Error (RTE) measures the differences between the ground truth and the predicted translation vector. It is defined as follows:

$$\mathrm{RTE} = \|t_p - t_g\|_2, \tag{5}$$

where t_p is the predicted translation vector and t_g is the ground truth translation vector. Relative Rotation Error (RRE) measures the differences between the ground truth and the predicted rotation matrices. It is defined as follows:

$$\mathrm{RRE} = \arccos\left(\frac{\mathrm{tr}(R_p^T \cdot R_{gt} - 1)}{2}\right), \tag{6}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. R_p and R_{gt} are the predicted rotation matrix and ground truth rotation matrix, respectively.

3 Additional Experiments

3.1 Evaluation on Outdoor KITTI Dataset

Table 1: Registration evaluation results on KITTI dataset under strict registration threshold (RRE<0.5° and RTE<20cm), with the top and second-ranking results highlighted in bold and underlined, respectively.

Method	RTE(cm)	RRE(°)	RR(%)
Predator [3]	5.8	0.19	93.1
GeoTransformer [5]	5.5	<u>0.18</u>	94.2
Lepard [4]	<u>5.2</u>	0.21	<u>94.5</u>
SemReg (ours)	**5.1**	**0.13**	**95.9**

We follow [4] and add an experiment on the outdoor KITTI [2] dataset to verify its robustness. Since common practices [3–5] show saturated registration performance (RR>99%) under the default threshold (RRE<5° and RTE<2m), we show the superior results under strict registration thresholds (RRE<0.5° and RTE<20cm) in Table 1.

Table 2: Registration evaluation results on ModelNet dataset, with the top and second-ranking results highlighted in bold and underlined, respectively.

Method	ModelNet40			ModelLoNet40		
	RRE(cm)	RTE(°)	CD(10^{-3})	RRE(cm)	RTE(°)	CD(10^{-3})
Predator [12]	1.739	0.019	0.89	5.235	0.132	8.3
Lepard [15]	1.532	<u>0.017</u>	0.81	4.854	0.104	4.6
GeoTransformer [22]	<u>1.489</u>	0.021	<u>0.71</u>	<u>3.968</u>	<u>0.099</u>	<u>3.3</u>
SemReg (ours)	**1.466**	**0.012**	**0.69**	**3.112**	**0.079**	**3.1**

3.2 Evaluation on Non-RGBD Data

Our proposed SemReg is designed for RGB-D data where the corresponding aligned 2D image is required to extract semantic features. Nevertheless, in this experiment, we show promising extensibility of our method to the case of point clouds without RGB-D data. We evaluate our SemReg on ModelNet40 [6]. It comprises 12,311 CAD models of man-made objects from 40 different categories. Following [3], we use 5,112 samples for training, 1,202 samples for validation, and 1,266 samples for testing. Since ModelNet40 does not contain RGB information, we render multi-view 2D images by projecting the input point cloud from three orthogonal views (along the x, y, and z axes) and incorporating those depth maps as input 2D images. We follow [3, 5] and use RRE/RTE/CD for a fair comparison. The experimental results are shown in Table 2. It shows our method's applicability to cases without images and its superiority over recent methods on both ModelNet40 and ModelLoNet40.

3.3 Influence of SCM Threshold τ_m

We assess the performance of our approach across various thresholds (τ_m) within the Semantics Constrained Matching (SCM) module. Detailed experimental findings are presented in Table 3.

Table 3: Evaluation results on 3DMatch and 3DLoMatch, with the top results highlighted in bold. The selected value is marked with $*$.

	3DMatch			3DLoMatch		
	FMR	IR	RR	FMR	IR	RR
$\tau_m = 0.1$	98.5	55.4	94.5	86.4	25.8	74.9
$\tau_m = 0.3*$	**98.6**	56.2	**94.7**	**88.4**	29.4	**75.9**
$\tau_m = 0.5$	98.4	**56.4**	94.3	88.3	**29.5**	75.3

The threshold τ_m in the Semantics Constrained Matching (SCM) module is used to suppress the potential erroneous matching outside of semantic correlated regions. Therefore, when we set τ_m to a lower value, FMR, IR, and RR saw marginal decreases. However, if τ_m is too high (e.g. 0.5), it inevitably eliminates

some borderline correspondences that have positive impacts during RANSAC, leading to incorrect registrations. We therefore select $\tau_m = 0.3$ as the default setting for all experiments.

3.4 Influence of GMSP Threshold η

The experimental results are shown in Table 4. The threshold η in the Gaussian Mixture Semantic Prior (GMSP) module is used to select the semantically salient points.

Table 4: Evaluation results on 3DMatch and 3DLoMatch, with the top results highlighted in bold. The selected value is marked with $*$.

	3DMatch			3DLoMatch		
	FMR	IR	RR	FMR	IR	RR
$\eta = -0.3$	98.1	56.1	94.3	87.9	29.3	75.1
$\eta = -0.1$	98.6	56.0	94.4	88.3	29.1	75.6
$\eta = 0.1*$	**98.6**	**56.2**	**94.7**	**88.4**	**29.4**	**75.9**
$\eta = 0.3$	98.4	55.9	93.8	87.4	28.6	74.2
$\eta = 0.5$	97.1	55.5	93.6	87.2	28.6	74.3

We observe that when η is too low, it slightly affects the RR. This is because a low threshold η might induce more points in semantically non-related regions to be selected as salient points, which is harmful to learning robust point features. On the contrary, when η is too high, RR saw a significant drop for both 3DMatch and 3DLoMatch. We suspect that a high threshold η might erroneously exclude some semantically similar points and thus lead to subpar feature learning. We therefore select $\eta = 0.1$ as the default setting for all experiments. Please refer to Figure 4 in the main paper for the visualization of the impact of η.

3.5 Number of SGFI Blocks

We tested various numbers of Semantics Guided Feature Interaction (SGFI) layers and the experimental results are shown in Table 5. We observe a significant decrease in performance when the model has only one SGFI layer. For RR results, the model with two SGFI layers achieved the best results on both 3DMatch and 3DLoMatch. Adding extra layers to the model does not lead to a better RR performance, which, however, increases the training time and the computation memory. We therefore select the model with two SGFI layers as the default setting for all experiments.

3.6 2D Semantic Backbone

We show the influence of the employed 2D semantic backbone. We use the pre-trained DINOV1-ResNet50 as the weaker 2D semantic backbone for comparison. Table 6 shows using a weaker backbone leads to a performance decline across metrics.

Table 5: Evaluation results on 3DMatch and 3DLoMatch, with the top results highlighted in bold. The selected value is marked with $*$.

number of SFGI	3DMatch FMR	3DMatch IR	3DMatch RR	3DLoMatch FMR	3DLoMatch IR	3DLoMatch RR
1	96.4	54.8	93.1	85.1	25.9	71.1
2*	**98.6**	**56.2**	**94.7**	88.4	**29.4**	**75.9**
3	98.2	55.8	94.5	**88.7**	29.2	75.4
4	98.5	56.0	**94.7**	88.1	29.3	75.8

Table 6: Evaluation results on 3DMatch and 3DLoMatch with different 2D semantic backbones.

Method	3DMatch FMR	3DMatch IR	3DMatch **RR**	3DLoMatch FMR	3DLoMatch IR	3DLoMatch **RR**
w. ResNet50	97.8	55.4	92.4	86.4	26.9	71.2
w. Default	**98.6**	**56.2**	**94.7**	**89.8**	**29.4**	**75.9**

3.7 Borderline Correspondences

We increase the distance threshold d from the default $0.1m$ to $0.15m$ and $0.2m$. Table 7 shows our method yields higher IR and more borderline correspondences (slightly above d) than other methods. These borderline correspondences contribute to registration [4], leading to superior RR of our method.

Table 7: Evaluation results on inlier ratio with different thresholds.

Method	3DMatch $0.1m$	3DMatch **0.15m**	3DMatch **0.2m**	3DLoMatch $0.1m$	3DLoMatch **0.15m**	3DLoMatch **0.2m**
Predator [5]	58.0	67.7(↑9.7)	72.0(↑14.0)	26.7	33.0(↑6.3)	37.5(↑10.8)
Lepard [4]	55.5	66.2(↑10.7)	72.6(↑17.1)	26.0	35.4(↑9.4)	39.2(↑13.2)
Ours	56.2	**72.8**(↑**16.6**)	**82.0**(↑**25.8**)	29.4	**43.4**(↑**14.0**)	**51.1**(↑**21.7**)

4 Analysis

4.1 Low Overlap Cases

We analyze the effectiveness of our method on low overlap cases. Figure 2 shows an extremely low overlap of 12.53% from the 3DLoMatch dataset. GeoTransformer [5] failed to extract reliable correspondences for registration due to the lack of geometric features (figure 2 (c)). However, since the overlap has useful semantics, i.e., a chair is included in the overlapping area, our method is able to achieve accurate registration in such a harsh case. Concretely, given a query point in the source point cloud (red dot in figure 2(a)), our method successfully identifies the semantic region in the target (red area in figure 2(b)) and achieves successful registration (figure 2(d)).

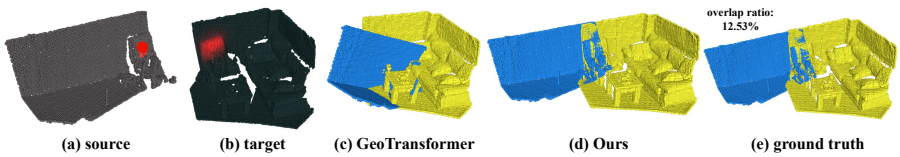

Fig. 2: Visualization on an extremely low overlap case.

4.2 Inference Speed

Table 8 shows the comparisons of the time consumption in the inference stage. Our method ranks second and is faster than SpinNet [1] and Predator [3].

Table 8: Inference speed comparisons.

Method	SpinNet [1]	Predator [3]	Lepard [4]	Ours
Time (ms)	4832	184	**142**	<u>166</u>

4.3 Failure Cases

Figure 3 shows a failure case. In this case, the overlap between two point clouds is extremely low. Meanwhile, the overlapping area has ineffective semantics, i.e., the overlap mainly involves a flat area that provides limited semantics.

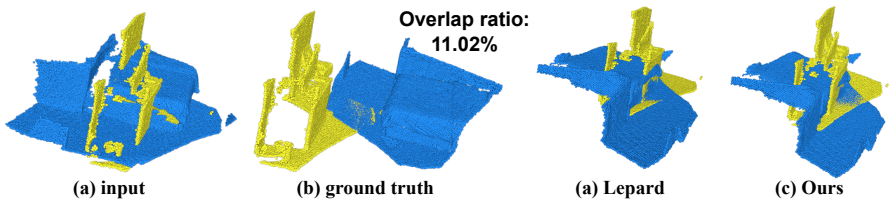

Fig. 3: Visualization on a failed case with low overlap and limited semantics.

5 Erroneous RGB-D data in 3DMatch

Our proposed method is a cross-modal approach and we incorporate the RGB images provided by the 3DMatch dataset. Each point cloud instance is generated from 50 RGB-D frames. However, some point cloud samples presented in the dataset are partially cropped from the original point cloud, which was generated from RGB-D frames, as shown in Figure 4. As such, we first compute the

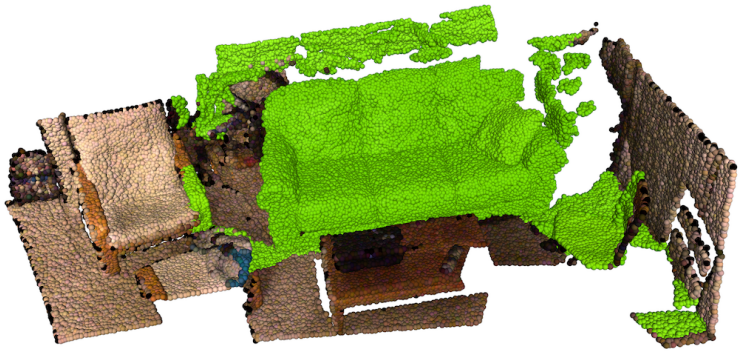

Fig. 4: An example in the 3DMatch dataset where the provided point cloud (green) is not fully overlapped with the point cloud (other colors) generated with the provided RGB-D frames.

overlapping ratio between the RGB-D frames and the point cloud sample, and then evenly sample N images from the frames with an overlapping ratio higher than 30%.

6 Visual Results

We provide more visual results on the 3DMatch and 3DLoMatch datasets. For the 3DMatch dataset, we compare the qualitative results from Lepard [4] and GeoTransformer [5] in Figure 5. We observe that our method still performs quite well in cases where there is less geometrical distinctiveness (e.g. 1^{st}, 2^{nd}, and 5^{th}), thanks to our robust semantic features obtained from corresponding 2D images.

For the 3DLoMatch dataset, we compare our method with the qualitative results from Lepard [4] in Figure 6. In addition to the registration results, we also show the visualization of the features with T-SNE and the putative correspondences. We observe from the feature visualization that even in the low-overlap cases, our proposed SemReg is able to extract consistent and distinctive features, leading to successful registration.

References

1. Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y.: Spinnet: Learning a general surface descriptor for 3d point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11753–11762 (2021)
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
3. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4267–4276 (2021)

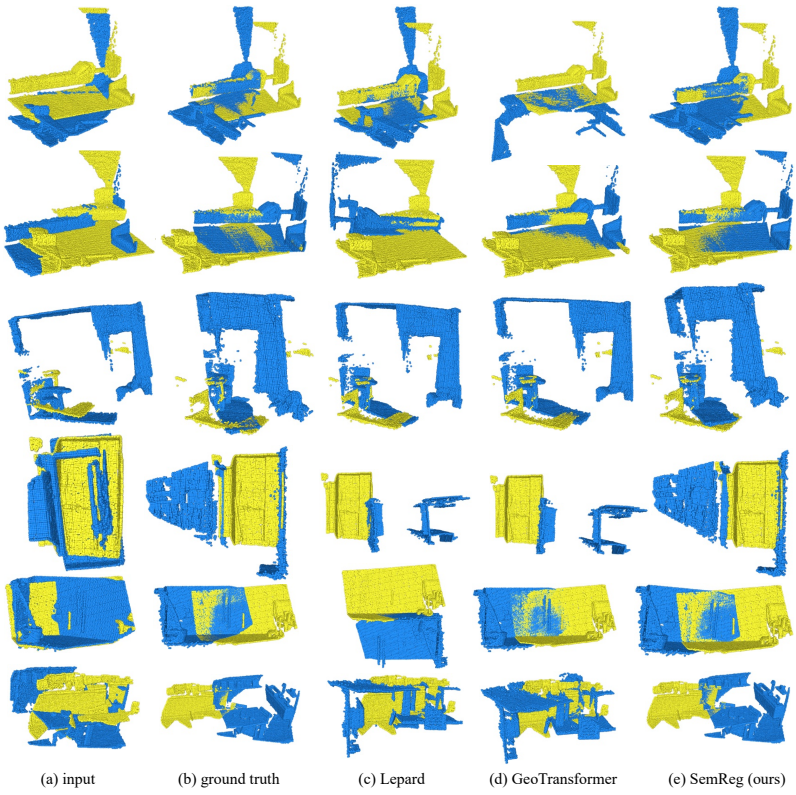

Fig. 5: Comparisons of visual registration results.

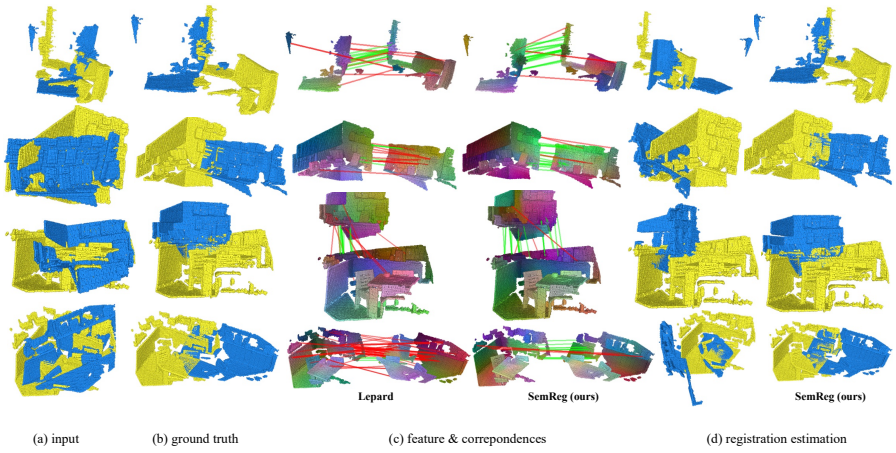

Fig. 6: Visual point cloud matching and registration results on 3DLoMatch.

4. Li, Y., Harada, T.: Lepard: Learning partial point cloud matching in rigid and deformable scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5554–5564 (2022)
5. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11143–11152 (2022)
6. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
7. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1802–1811 (2017)