Fully-geometric Cross-attention for Point Cloud Registration Supplementary Materials

Anonymous 3DV submission

Paper ID 213

In this supplementary file, we first introduce the experimental details in Sec. 1, including implementation details, correspondence sampling, and model architecture. We then report more results in Sec. 2.

1. Experimental Details

1.1. Implementation details

Running details. We utilized PyTorch to implement our method and trained it on a system consisting of one Quadro GV100 GPU (32G) and two Intel(R) Xeon(R) Gold 6226 CPUs. Our training process consisted of 50 epochs for 3DMatch and 90 epochs for KITTI, with a batch size of 1. We used the AdamW optimizer with a weight decay of 1e-6. The initial learning rate was set to 1e-4 for both datasets, but it was decreased by a factor of 0.05 after each epoch on 3DMatch and every 4 epochs on KITTI.

The encoder and decoder architectures used were identical to those in [9]. For training, we randomly selected 128 ground-truth super-point correspondences, while for testing, we used 256 putative matches. For the geometric Transformer, we repeated it 3 times. For Fine matching on 3DMatch and KITTI, we first sampled 64 points for each patch, then performed geometric self-attention in each patch to produce more distinctive descriptors for correspondence prediction. On 3DCSR, we generated 48 points for each patch if a patch contains more than 48 points, then performed geometric self-attention.

Correspondence sampling. Our approach for sampling various numbers of interest points is based on CoFiNet [9]. To obtain correspondences, we use a probability sampling method that considers the product of the confidence scores for both coarse and fine matching, i.e., $\bar{\Gamma} * \Gamma$.

Architecture. Our approach utilizes an encoder-decoder framework that employs KPConv operations. We have incorporated two attention-based networks, which are geometry-enhanced, to facilitate context aggregation and geometric embedding. For further information regarding our network architecture, please refer to Fig. 1.

Table 1. Computation time analysis on both 3DMatch and 3DLoMatch datasets.

	RR		Time (s)		
Method	3DM	3DLM	Model	Pose	Total
FCGF [4]	85.1	40.1	0.052	3.326	3.378
D3Feat [2]	81.6	37.2	0.024	3.088	3.112
SpinNet [1]	88.6	59.8	60.248	0.388	60.636
Predator [5]	89.0	59.8	**0.032**	5.120	5.152
CoFiNet [9]	89.3	67.5	0.115	1.807	1.922
GeoTrans [6]	92.0	75.0	0.075	1.558	**1.633**
FLAT (ours)	**92.4**	**78.6**	0.412	**1.502**	1.914

2. Additional results

Computation time analysis. We computed the average inference time of our proposed method and compared it to that of the baseline methods on 3DMatch and 3DLoMatch. It is worth mentioning that each method consists of two stages, which are feature or correspondence extraction and transformation recovery using RANSAC. We report the inference times for both stages. For baselines, we use the codes and pre-trained models provided by the authors and run them in our environment for a fair comparison. While our approach may be marginally slower than certain baselines in the correspondence prediction stage, it outperforms them in reliably extracting correspondences. All experiments were conducted on the 3DMatch testing set, using a single Tesla V100-PCIE GPU (32G) and two Intel(R) Xeon(R) Gold 6226 CPUs.

Test with very low overlap ratios. Certainly! Here's a revised version of your paragraph for improved clarity and coherence:

"We selectively analyzed scenarios within 3DMatch where the overlap ratio falls between 1.0% and 10%. As demonstrated in Tab. 2, the introduction of full geomet-

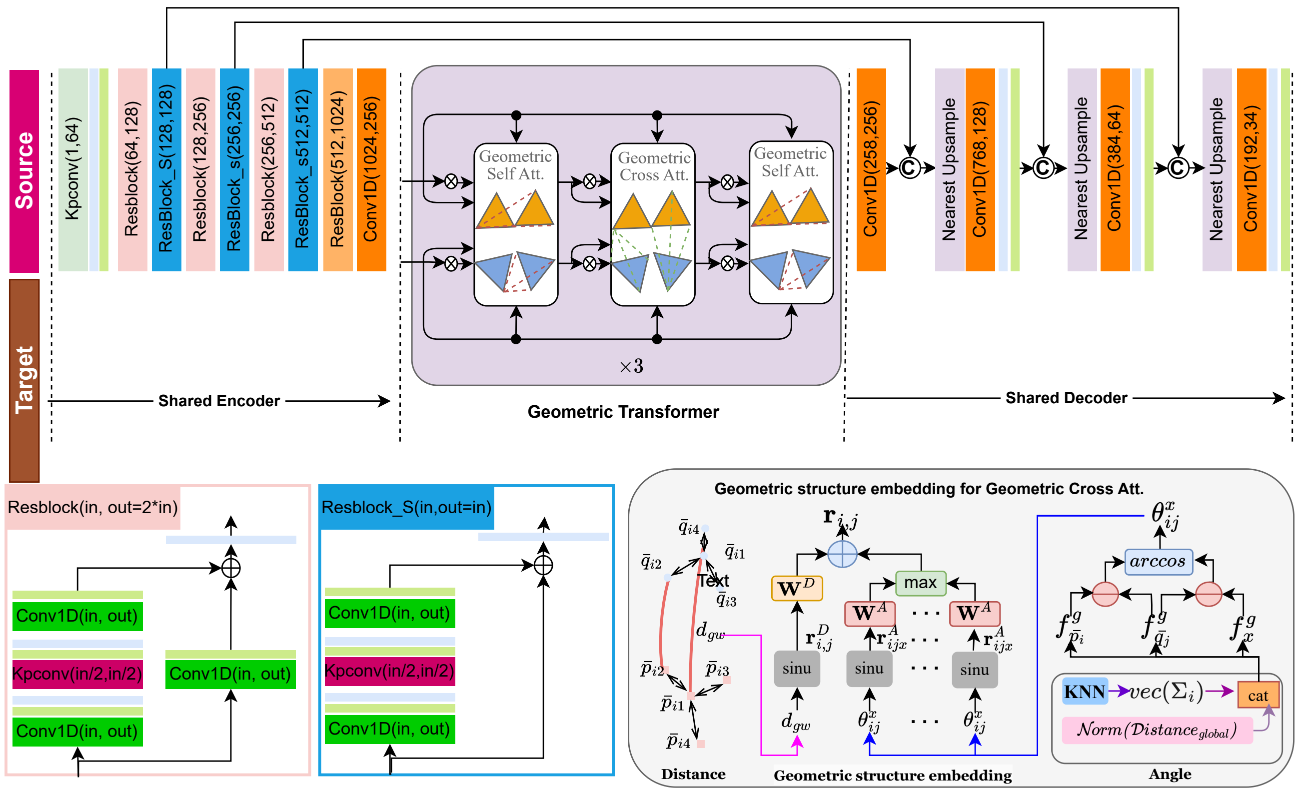

Figure 1. Detailed architecture FLAT. Within the self- and cross-attention modules, the multi-head attention part utilizes four heads. Angle and distance provide geometry information for cross-attention.

Table 2. Registration results on 3DMatch where the overlap ratio ranges between 1.0% to 10%.

	RR↑	FMR↑	IR↑
GeoTr	38.2	66.9	17.0
Ours	**46.9**	**76.3**	**21.4**

ric cross-attention in our approach significantly enhances performance over GeoTransformer. This advancement can be attributed to the primary distinction between our FLAT and GeoTransformer: the implementation of full geometric cross-attention, which effectively aids in identifying overlapping regions. Consequently, FLAT exhibits heightened proficiency in detecting more accurate correspondences in cases with low overlap.

Failure cases. Fig. 2 shows two failure cases on 3DLoMatch wherein the overlapping regions are planar surfaces, lacking geometric information. We analyze cases with low overlap ratios: 10.2% top case, 13.8% bottom case. Correctly matched points are colored in red, while incorrectly matched points are black. Although several features are correctly matched at the coarse level, the refinement stage produces uninformative features due to the ambiguous geometric structure of planar surfaces, failing registration.

Registration results on 3DMatch and 3DLoMatch. Following REGTR [8], we conducted further analysis of the Relative Rotation Errors (RRE) and Relative Translation Errors (RTE) to assess the accuracy of successful registrations. Tab. 3 presents the results of the various methods, with the best performance highlighted in bold and the second-best results underlined. Our method demonstrates superior performance on 3DLoMatch, achieving the lowest average rotation (RRE) and translation (RTE) errors across scenes. Additionally, our method exhibits the highest average registration recall, indicating the final performance on point cloud registration (92.4% on 3DMatch and 78.6% on 3DLoMatch).

Qualitative results of registration. Fig. 3 shows visual results on KITTI. The correspondences extracted by FLAT are used as input for RANSAC to estimate the relative transformation. These outcomes underscore the efficacy of our method in outdoor datasets. They demonstrate the adaptability and strong performance of the full geometric cross-attention mechanism inherent in FLAT, even within outdoor settings.

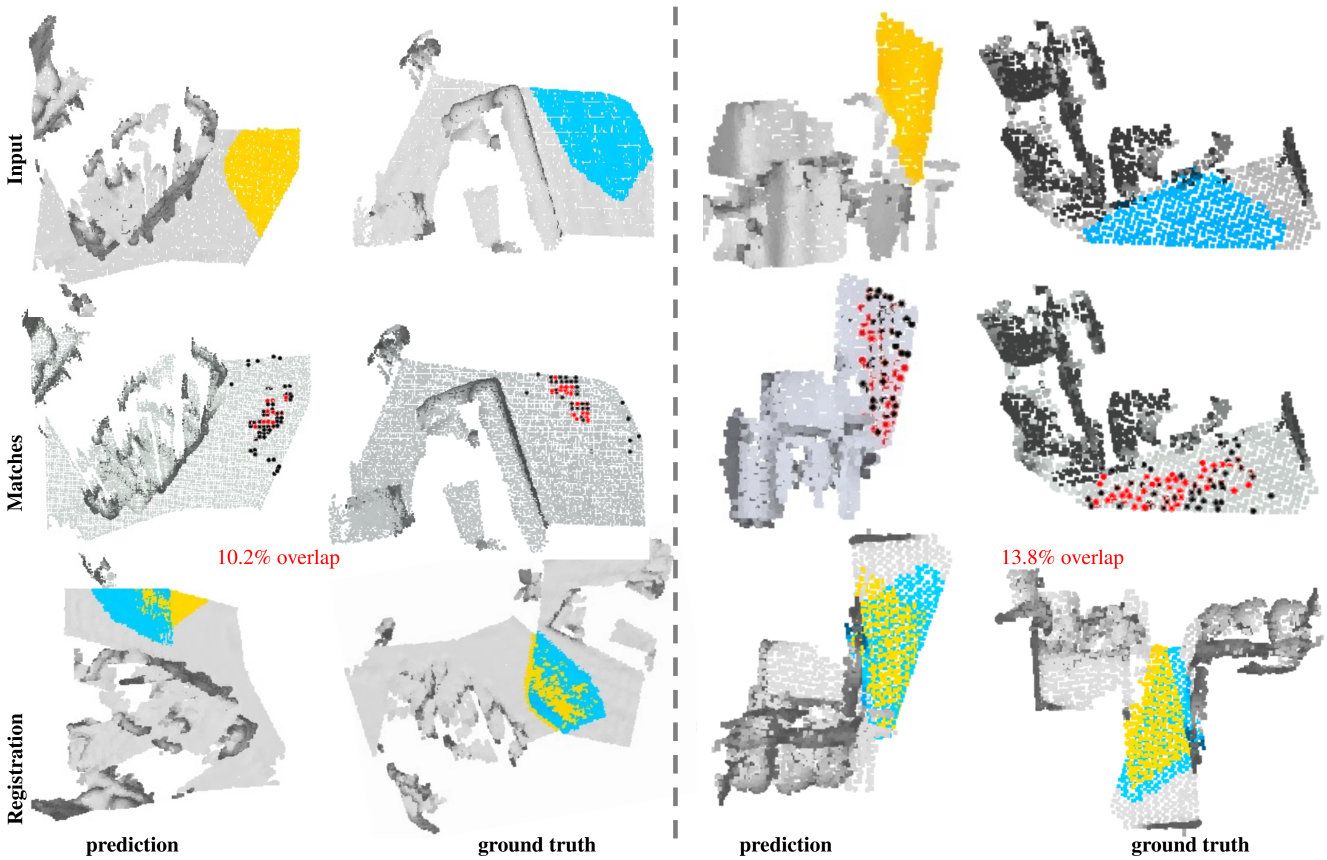

Figure 2. Example qualitative registration results for failure cases on 3DLoMatch.

Table 3. Results on both 3DMatch and 3DLoMatch datasets. The best results for each criterion are labeled in bold, and the second-best results are underlined.

Method	3DMatch RR↑	3DMatch RRE ↓	3DMatch RTE ↓	3DLoMatch RR ↑	3DLoMatch RRE ↓	3DLoMatch RTE ↓
FCGF [4]	85.1%	1.949	0.066	40.1%	3.147	0.100
D3Feat [2]	81.6%	2.161	0.067	37.2%	3.361	0.103
OMNet [7]	35.9%	4.166	0.105	8.4%	7.299	0.151
DGR [3]	85.3%	2.103	0.067	48.7%	3.954	0.113
Predator1K [5]	90.5%	2.062	0.068	62.5%	3.159	0.096
CoFiNet [9]	89.7%	2.147	0.067	67.2%	3.271	0.090
GeoTrans [6]	92.0%	1.808	0.063	74.0%	2.934	0.089
REGTR [8]	92.0%	**1.567**	**0.049**	64.8%	2.827	0.077
FLAT (ours)	**92.4%**	<u>1.690</u>	<u>0.053</u>	**78.6%**	**2.599**	**0.070**

Time Cost Comparison with GeoTransformer We have reported the time costs on Tab. 4 in the Appendix of the submitted materials. We also compare our method with RoITr and GeoTr in terms of time costs, as the table below shows.

The "model" is the time for feature extraction and correspondence search, while the "pose" is for transformation estimation. Our model time is indeed a bit heavier; this is mainly because the computation of the Gromov-Wasserstein distance is expensive.

Method	**Model (s)↓**	**Pose (s)↓**	**Total (s)↓**
RoITr	**0.053**	1.524	**1.577**
GeoTr	0.075	1.558	1.633
FLAT (*Ours*)	0.412	**1.502**	1.914

Table 4. Comparison of methods based on Model, Pose, and Total time in seconds.

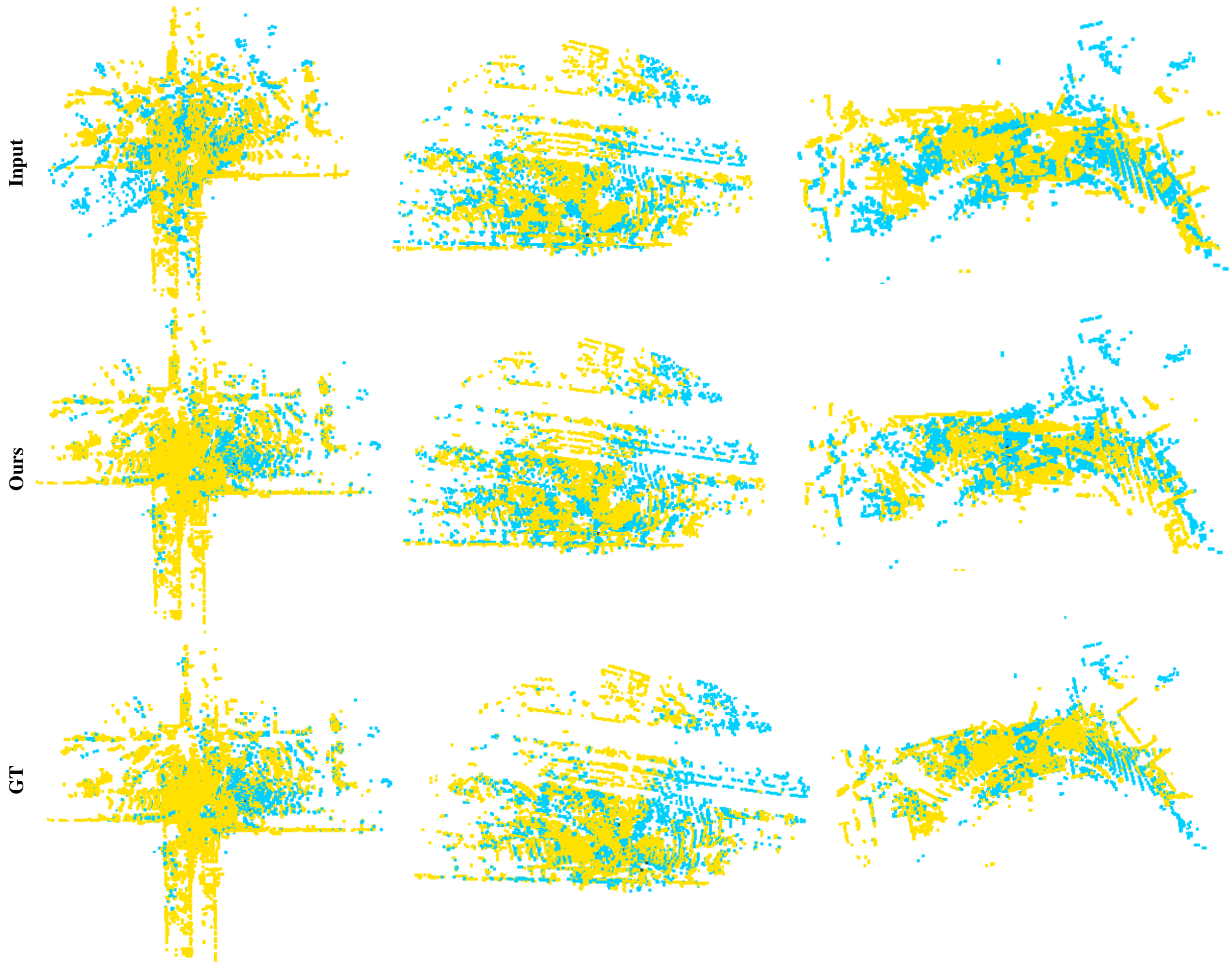

Figure 3. Example qualitative registration results on KITTI. The (Input) column exhibits the input point cloud pairs, the (Our) column demonstrates the estimated registration, and the (GT) column presents the ground truth alignment.

References

[1] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *CVPR*, pages 11753–11762, 2021. 1

[2] Xuyang Bai and et al. D3feat: Joint learning of dense detection and description of 3d local features. In *CVPR*, pages 6359–6367, 2020. 1, 3

[3] Christopher Choy and et al. Deep global registration. In *CVPR*, pages 2514–2523, 2020. 3

[4] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, pages 8958–8966, 2019. 1, 3

[5] Shengyu Huang and et al. Predator: Registration of 3d point clouds with low overlap. In *CVPR*, pages 4267–4276, 2021. 1, 3

[6] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, pages 11143–11152, 2022. 1, 3

[7] Hao Xu, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In *ICCV*, pages 3132–3141, 2021. 3

[8] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *CVPR*, pages 6677–6686, 2022. 2, 3

[9] Hao Yu and et al. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *NeurIPS*, 34, 2021. 1, 3