ColabSfM: Collaborative Structure-from-Motion by Point Cloud Registration

Supplementary Material

In this supplementary material, we provide details on the training, RANSAC implementation, computation of point cloud overlap, additional registration results, and additional qualitative results that could not fit in the main paper.

1. Training Details

The models were trained on a NVIDIA 4090 GPU, both in 3DMatch [10] and our dataset. We used the Adam Optimizer, trained for 150 epochs, with a batch size of 1, learning rate of 0.0001, and exponential decay 0.05.

2. RANSAC

We solve for the pose using the RANSAC [1] implementation from Open3D [11] on the matches retrieved by the neural networks. The samples consist of three points, and inlier counting is performed by checking if the correspondence distance is below $\tau_{\text{IR}} = 0.05$. We use at most 1000 correspondences, if the models return more than 1000 matches, we random sample 1000 matches. If correspondence scores are given by the methods, we use them to weight the correspondence sampling. Rotation and translation error are computed as

$$\epsilon_R(R) = \text{acos}\left(\frac{\text{trace}\left(R^{-1}R_{\text{GT}}\right)-1}{2}\right) \tag{1}$$

$$\epsilon_{\mathbf{t}}(\mathbf{t}) = \|\mathbf{t} - \mathbf{t}_{\text{GT}}\|, \tag{2}$$

respectively. Where $\|\cdot\|$ represents the Euclidean norm.

3. Computing Overlaps

We compute the overlap between two point clouds as the approximate intersection over union of the points. We define the directional intersection of $\mathbf{P} \rightarrow \mathbf{Q}$ as

$$\text{I}_{\mathbf{P}\rightarrow\mathbf{Q}} = \sum_{p\in\mathbf{P}} (\min_{q\in\mathbf{Q}} \text{dist}(Rp + t, q) \leq \tau), \tag{3}$$

with $\tau = 0.1$, and the union as

$$\text{U}_{\mathbf{P}\rightarrow\mathbf{Q}} = |\mathbf{P}|. \tag{4}$$

The directional IoU from $\mathbf{P} \rightarrow \mathbf{Q}$ is then

$$\text{IoU}_{\mathbf{P}\rightarrow\mathbf{Q}} = \frac{\text{I}_{\mathbf{P}\rightarrow\mathbf{Q}}}{\text{U}_{\mathbf{P}\rightarrow\mathbf{Q}}}. \tag{5}$$

We define the overlap between the two point clouds as the geometric mean of their directional IoUs, *i.e.*,

$$\text{IoU}_{\mathbf{P},\mathbf{Q}} = \sqrt{\text{IoU}_{\mathbf{P}\rightarrow\mathbf{Q}} \cdot \text{IoU}_{\mathbf{Q}\rightarrow\mathbf{P}}}. \tag{6}$$

4. Cambridge Landmarks Additional Results

The rotation and translation errors computed using (1) and (2) are presented in Tab. 1. The methods trained only on 3DMatch [10] present poor results, which is expected given the low number of inliers (*cf*. **??** of the main paper). The methods trained and/or fine-tuned on our dataset have small errors, with the performance being similar for all four. To assess the performance of the registration methods against visual cues, we reconstructed all scenes using SIFT descriptors, subsampled the 3D reconstructions to 30k points (the same number used for the registration models) and computed the average descriptor for each 2D point in a 3D point's track. Then, we used Nearest Neighboor (NN) matching and RANSAC to find the pose, this method is called SIFT+NN, see Tab. 1. Notice that for most scenes, our descriptor-free approach has comparable results to the descriptor-based approach, validating our results. The most challenging scene, for the registration methods, was *Old Hospital*, which consists of a facade with repetitive structure. The point cloud consists mostly of the facade and further way points in much sparser areas leading to a set of bad matches found in those regions, see Fig. 3. The same happens in *Shop Facade* which consists of a corner shop, which also contains a high amount of points in further and sparser areas, resulting in a high number of bad matches. Nevertheless, the correct matches found by the fine-tuned versions RefineRoITr and RoITr [9] are enough to find accurate poses for the registration.

We additionally further investigate out-of-distribution results of our model trained on SOSNet and SIFT reconstructions using DISK [7] and LightGlue [4]. Results are presented in Tab. 2. Our trained models demonstrate the ability to generalize to reconstructions of a different nature, however showing a slight decrease in performance. It is likely that including a more diverse set of detectors in the training set would decrease this gap, and further improve the generalizability of our approach.

5. Additional qualitative results

SfM Registration Benchmark: We evaluate RefineRoITr trained from scratch on our proposed dataset against OverlapPredator [2] and RoITr [9] trained on 3DMatch [10] qualitatively. The results are presented in Fig. 1. We present two more pairs for two different test scenes, namely the Brandenburger Tor and Piazza San Marco test scenes. From inspection, we can see that OverlapPredator tends to produce a high number of matches, but only a very small fraction are inliers (0.8% on aver-

Table 1. **Results of SfM registration on Cambridge Landmarks [3]**. We evaluate unknown relative orientation SO(3) + unknown relative position = SE(3). The top portion contains methods only trained on the 3DMatch (3DM) dataset, the middle portion methods trained only on our proposed dataset (Mega), while the lower portion contains methods trained on the former and fine-tuned on the latter. Rotation error is reported in degrees, translation error is unitless since the pointclouds are scaled.

	Great Court		Kings College		Old Hospital		Shop Facade		St Mary's Church	
Method	ϵ_R	$\epsilon_\mathbf{t}$	ϵ_R	$\epsilon_\mathbf{t}$	ϵ_R	$\epsilon_\mathbf{t}$	ϵ_R	$\epsilon_\mathbf{t}$	ϵ_R	$\epsilon_\mathbf{t}$
SIFT + NN	**0.1°**	0.09	**0.2°**	0.02	**0.2°**	0.05	**0.1°**	**0.01**	**0.04°**	**0.01**
OverlapPredator [2] (3DM)	6.9°	0.43	27.2°	0.24	135.5°	2.14	179.4°	0.82	170.7°	4.27
GeoTransformer [6] (3DM)	177.2°	7.26	81.2°	15.13	179.0°	14.65	123.3°	19.27	146.8°	12.13
PareNet [8] (3DM)	173.7°	4.24	149.4°	4.82	171.1°	4.69	32.74°	1.10	176.07°	3.48
RoITr [9] (3DM)	179.4°	6.93	157.3°	22.42	76.7°	0.0	176.9°	1.09	99.3°	2.14
RefineRoITr (3DM)	73.8°	1.76	90.3°	6.13	81.6°	3.97	84.8°	1.04	138.6°	15.83
RefineRoITr (Mega)	0.5°	**0.02**	0.5°	0.02	1.16°	0.05	0.9°	0.02	0.3°	**0.01**
RefineRoITr (Mega w\ color)	0.3°	**0.02**	0.6°	0.03	3.0°	0.15	0.5°	0.03	0.3°	**0.01**
RoITr [9] (3DM + Mega)	0.4°	**0.02**	0.6°	**0.01**	2.6°	0.10	2.6°	0.02	0.2°	0.02
RefineRoITr (3DM + Mega)	0.5°	0.03	**0.2°**	0.02	2.1°	**0.04**	0.9°	0.04	0.3°	**0.01**

Table 2. **Results of Out-of-Distribution SfM registration on Cambridge Landmarks [3]**. Here we take a network trained on SIFT and SOSNet reconstructions, and evaluate it on DISK reconstructions. We evaluate unknown relative orientation SO(3) + unknown relative position = SE(3). The top portion contains methods only trained on the 3DMatch (3DM) dataset, the middle portion methods trained only on our proposed dataset (Mega), while the lower portion contains methods trained on the former and fine-tuned on the latter.

	Great Court		Kings College		Old Hospital		Shop Facade		St Mary's Church	
Method	IR	Matches	IR	Matches	IR	Matches	IR	Matches	IR	Matches
OverlapPredator [2] (3DM)	2.6	352	1.1	361	0	286	1.7	359	0	373
GeoTransformer [6] (3DM)	0	256	0	283	0	238	0	278	0	309
RoITr [9] (3DM)	0	400	0	733	0	256	0	**930**	0.2	867
RefineRoITr (3DM)	0	306	0	581	0	256	0	742	2.5	603
RefineRoITr (Mega)	36.9	**3942**	**54.1**	**2466**	9.5	**1437**	**13.5**	370	60.0	**4594**
RoITr [9] (3DM + Mega)	29.4	361	30.5	511	5.9	152	2.7	449	57.9	1528
RefineRoITr (3DM + Mega)	**39.8**	2848	49.5	1744	**9.7**	1152	10.1	485	**61.0**	4281

age). This results in poor pose estimations as can be seen in Fig. 1 second column bottom three rows. On the other hand RoITr yields fewer matches, but the majority of those are still not good, see the inlier ratios presented in the figure. Similar to OverlapPredator the registration results are not accurate enough to produce a good merging of the source and target point clouds, in yellow and blue respectively. Finally, our RefineRoITr is capable of finding a high number of matches, with the majority of them being inliers.

Quad6k [5]: To evaluate the generalization of RefineRoITr, we evaluate it and two baselines OverlapPredator [2] and RoITr [9] trained on 3DMatch [10]. The results are presented in Fig. 2. This dataset is more challenging and it consists mainly of a square surrounded by the facade of several buildings. The performance of the baselines is similar to what was observed in our SfM Registration Benchmark test scenes. When looking at the performance of RefineRoITr, we can see that when the pairs refer to distinctive structures, like the tower in both the first and third pairs of Fig. 2. The model is capable of finding good matches and find an accurate pose to register the point clouds. However, it struggles to find matches for the second pair, which contains the facade of two buildings with repetitive and symmetric structure, leading to a failure in the registration.

Cambridge Landmarks [3]: Additionally, we also present qualitative results of RefineRoITr trained only on our SfM Registration dataset in the five test scenes of the Cambridge Landmarks dataset. The matches and registration results are presented in Fig. 3. Our method obtained a high number of good matches and hence the accuracy of the registration results, see Tab. 1. However, it produces a number of outliers in sparser areas of the point clouds, which was not the case

in the other datasets see Figs. 1 and 2. Since the scenes in this dataset consist of videos instead of random sets of images, there is less viewpoint variability, which also leads to more far points being matched across multiple images and accepted by the SfM pipeline. The scenes where this was more evident were *Old Hospital* and *Shop Facade*, which present almost flat facades with repetitive and symmetric structure. This is a limitation of the current method, which indicates an avenue for future research.

6. Registration for Partial Scenes Reconstructed from Scratch

In practice, registering two reconstructions will face issues of drift, as the partial scenes will be reconstructed from scratch, and not retriangulated (as in our training data). To investigate the performance of our approach on this more realistic setting, we evaluated the SfM registration error on scenes from the Cambridge Landmarks dataset reconstructed from scratch (i.e., without using the ground truth poses for triangulation for one of the partial reconstructions), in Table 3. Encouragingly, we find that the registration error is still low, indicating that there is no major distribution shift between the two tasks.

7. Details on RoITr Architecture

Here we expand on our baseline RoITr's [9] architecture in more detail.

Encoder e_θ: The encoder consists of an Attentional Abstraction Layer (AAL) followed by e PPF Attention Layers (PAL) [9, Figure 2, Section 3.2].

Global Transformer g_θ: The global Transformer g_θ consists of g blocks, each consisting of a Geometry-Aware Self-Attention Module (GSM), followed by a Position-Aware Cross-Attention Module (PCM) [9, Section 3.3, Figure 2, Figure 5].

Decoder d_θ: The decoder d_θ consist of a Transition Up Layer (TUL) for upsampling and context aggregation, followed by d PALs [9, Section 3.2, Figure 2].

References

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[2] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4267–4276, 2021. 1, 2, 4, 5

[3] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Int. Conf. Comput. Vis.*, pages 2938–2946, 2015. 2, 6

[4] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *Int. Conf. Comput. Vis.*, 2023. 1

[5] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60:91–110, 2004. 2, 5

[6] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11143–11152, 2022. 2

[7] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Adv. Neural Inform. Process. Syst.*, 33:14254–14265, 2020. 1

[8] Runzhao Yao, Shaoyi Du, Wenting Cui, Canhui Tang, and Chengwu Yang. Pare-net: Position-aware rotation-equivariant networks for robust point cloud registration. In *ECCV*, 2024. 2

[9] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5384–5393, 2023. 1, 2, 3, 4, 5

[10] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1802–1811, 2017. 1, 2

[11] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 1

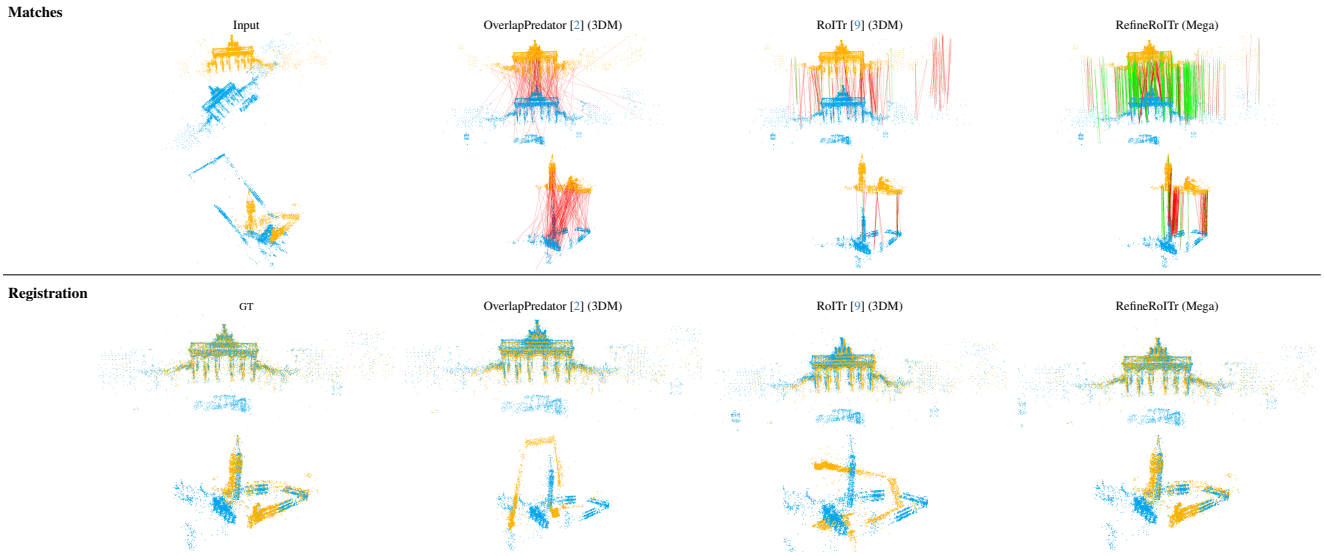

Figure 1. **Additional qualitative comparison on our dataset.** We compare our approach to previous point cloud registration methods on the Brandenburger Tor and Piazza San Marco test scenes (first and second row, respectively). Without training on our proposed SfM registration dataset (columns 2,3), previous methods are unable to produce sufficiently good matches (top two rows) to produce accurate relative pose estimation results (bottom two rows). In contrast, our proposed model RefineRoITr, trained on the proposed dataset, is able to register the scenes well.

Table 3. **Results of SfM registration on Cambridge Landmarks**. Results for point clouds reconstructed from scratch (*i.e.*, not retrianguled).

	Great Court		Kings College		Old Hospital		Shop Facade		St Mary's Church	
Method	ϵ_R	$\epsilon_{\mathbf{t}}$	ϵ_R	$\epsilon_{\mathbf{t}}$	ϵ_R	$\epsilon_{\mathbf{t}}$	ϵ_R	$\epsilon_{\mathbf{t}}$	ϵ_R	$\epsilon_{\mathbf{t}}$
SIFT + NN	0.32	**0.09**	**0.3**	**0.02**	**1.37**	**0.02**	**0.92**	**0.06**	0.49	**0.01**
RefineRoITr (Mega)	**0.30**	0.10	0.55	0.05	1.41	**0.02**	1.03	0.15	**0.21**	0.02
RoiTr (Mega)	0.86	0.17	0.46	0.04	10.23	1.07	2.71	0.53	0.50	0.03

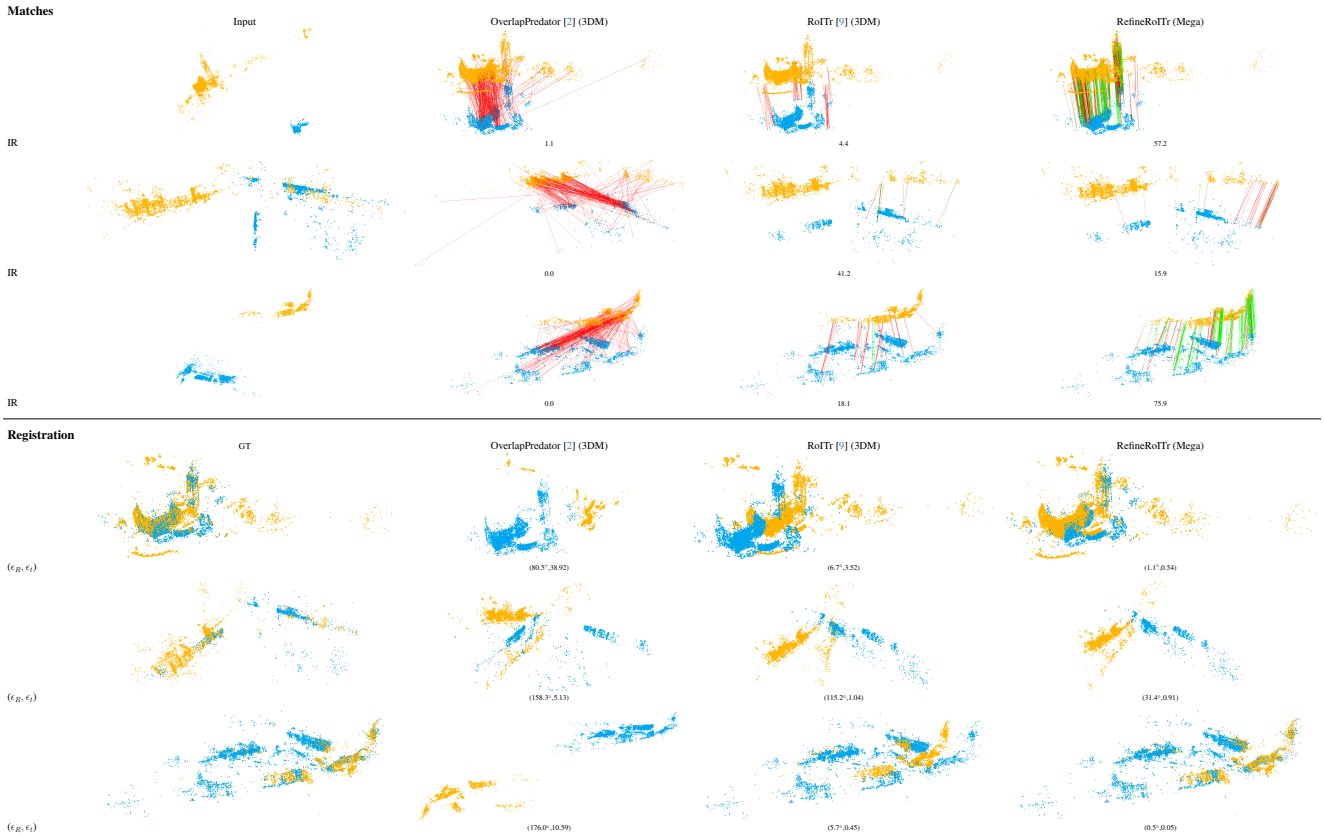

Figure 2. **Qualitative comparison Quad6k [5].** We compare our approach to previous point cloud registration methods on three test scenes (first to third row, respectively). Without training on our proposed SfM registration dataset (columns 2,3), previous methods are unable to produce sufficiently good matches (top three rows) and accurate relative pose estimation results (bottom three rows). In contrast, our proposed model RefineRoITr, trained on the proposed dataset, is able to find better matches and hence register the scenes well. The exception is the second scene, where our model struggles to find matches and fails to register the point clouds. The source and target point clouds are depicted in yellow and blue, respectively.

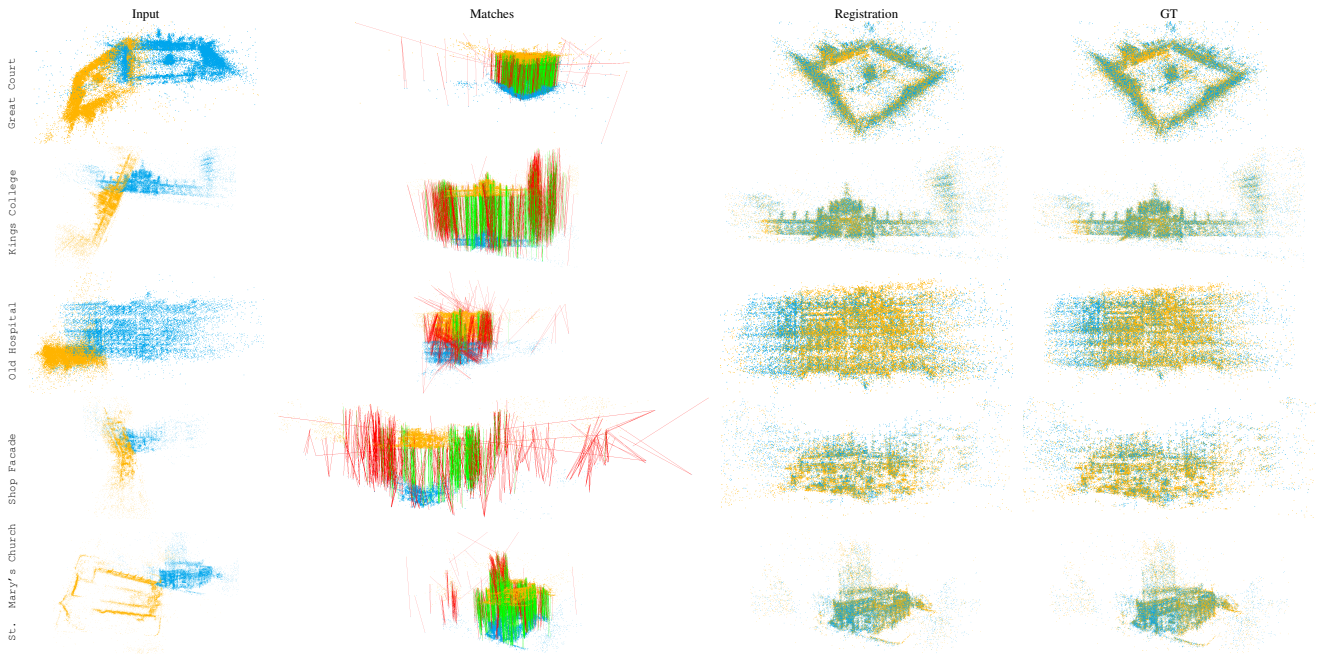

Figure 3. **Qualitative Results SfMRoITr on Cambridge Landmarks [3].** We evaluate SfMRoITr trained on our SfM Registration dataset on the five test scenes. We present the found matches and the results of the point cloud registration with the process presented in Sec. 2. The method is capable of obtaining accurate registration and a high number of good matches for all scenes. However, it tends to find outlier matches in sparser regions of the point clouds.