3D Local Features for Direct Pairwise Registration
Supplementary Material

Haowen Deng [†, *, ✣] Tolga Birdal [†, *] Slobodan Ilic [†, *]

[†] Technische Universität München, Germany, [*] Siemens AG, München, Germany
[✣] National University of Defense Technology, China

This document supplements our paper **3D Local Features for Direct Pairwise Registration**. It provides further quantitative ablation studies as well as qualitative results.

1. Ablation Study

Does multi-task training scheme help to boost the feature quality? In order to find out how multi-task training affects the quality of the learned intermediate features, we trained several networks with combinations of different supervision signals. For the sake of controlled experimentation, all networks are made to have the identical architecture. They are trained with the same data for 10 epochs. Hence, the only variable remains to be the objective function used for each group.

In total, there are four networks to be compared. The first one is trained with all the available supervision signals, i.e. reconstruction loss, feature consistency loss and pose prediction loss. Regarding the other three groups, each of the networks is trained with one of the three signals excluded. For simplicity, those groups are tagged as *All*, *No Reconstruction*, *No Consistency* and *No Pose* respectively. The fragment matching results using features from different networks are shown in Fig. 1.

As shown in Fig. 1, with all the training signals on, the learned features are the most robust and outperform all the others which lack at least one piece of information and thus suffer a performance drop. When no reconstruction loss is applied, the learned features almost always fail at matching. It is therefore the most critical loss to minimize. The absence of pose prediction loss has the least negative influence. Yet, it is necessary for RelativeNet to learn to predict the relative pose for given patch pairs. Without this the later stages of the pipeline such as hypotheses generation and verification cannot continue. These results validate that our multi-task training scheme takes full advantage of all the available information to drive the performance of learned local features to a higher level.

Invariant vs. -variant features in matching Our method extracts two kinds of local features using two different network components. The ones extracted by PPF-FoldNet are fully rotation-invariant, while local features of PC-FoldNet change as the pose of local patches vary. Experimentation contained in the paper used local features from PPF-FoldNet only to establish correspondences thanks to its superior property of invariance. Here, we use invariant and equivariant features to match fragment pairs separately, and compare their matching performance. This is important in validating our choice that invariant features are more suitable for nearest neighbor queries.

Fig. 2 exhibits the distribution of correspondence inlier ratio for the matched fragment pairs by using different local features. Matching results of equivariant features shows a huge amount of fragment pairs having correspondences with only a small fraction of inliers (less than 5%). Invariant features though, manage to provide many fragment pairs with a set of correspondences with over 10% true matches. It proves that invariant features are better at finding good correspondence set for further registration stage. All in all, rotation-invariant features extracted by PPF-FoldNet is more suitable for finding putative local matches. Note that this was also verified by [3].

Table 1. Average # of correspondences obtained by different methods of assignments. $K = k$ refers to retaining k-mutual neighbors.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	Closest
# Matches	335	1099	1834	2609	3664

More details for correspondence estimation methods In the main paper, we found out that a more relaxed condition for keeping neighbors lead to a better subsequent registration. However, this performance gain comes at a cost and hence introduces a trade-off. Tab. 1 tabulates the average number of putative matches found by different methods. As we can see, the size of correspondence set increases rapidly as we relax the standard and keep more neighbors. In return, this means more computation time in the following

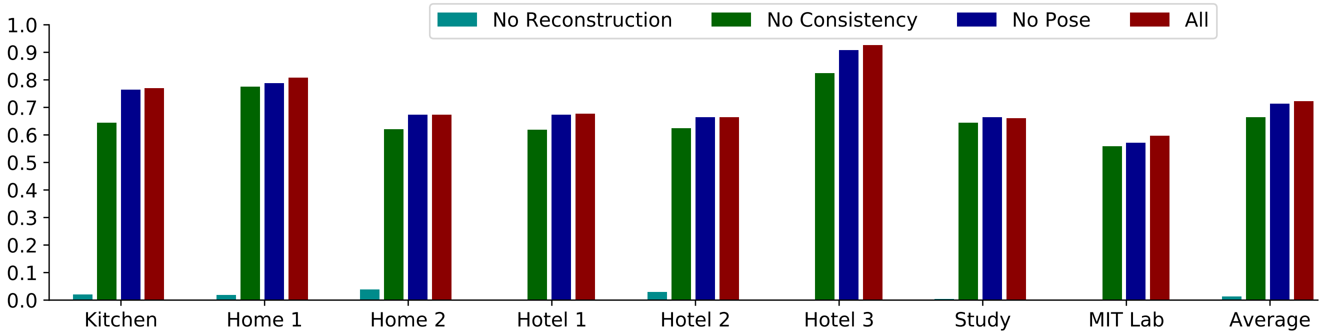


Figure 1. Influences of different supervision signals. Reconstruction is the most essential loss for our network to generate local features for matching tasks. Without it the descriptive-ness is lost. When all losses are combined, the network learns to extract the most powerful features and achieves the best performance.

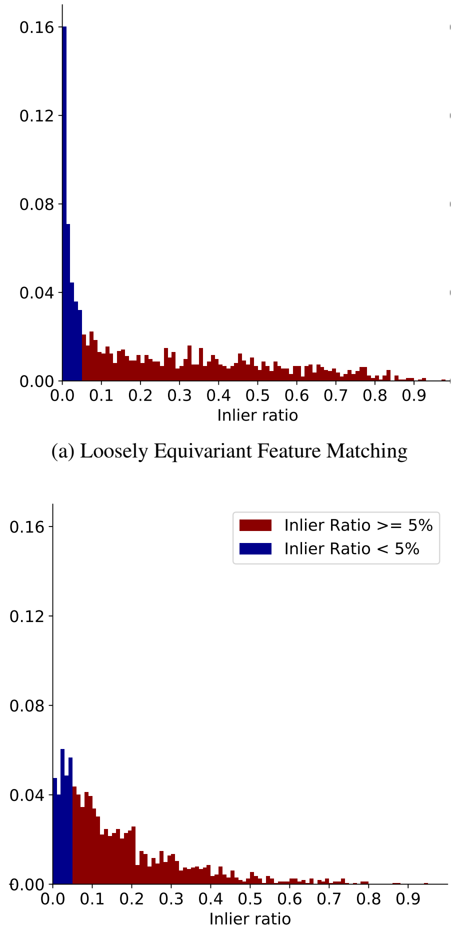
registration stage.

2. Quantitative Results

Distribution of hypotheses Fig. 3 shows the distribution of poses predicted by RelativeNet and poses determined by running RANSAC on the randomly selected subsets of corresponding points. Each hypothesis is composed of a rotational and translational part. The former is represented as a Rodrigues vector to keep it in $\mathbb{R}^3$. It is obvious that hypotheses predicted by RelativeNet are centered more around the ground truth pose, both in rotation and translation. It also reveals the reason why the hypotheses of our network could facilitate an easier and faster registration procedure.

Qualitative comparison against RANSAC Fig. 4 shows some challenging cases where only a small number of correct correspondences are established. In these examples, RANSAC fails to recover the pose information from the small set of inliers hidden in a big set of mismatches. However, a registration procedure with the aid of RelativeNet could succeed with a correct result. The qualitative comparison demonstrates that our method is robust at registering fragment pairs even in extreme cases where insufficient inliers are presented.

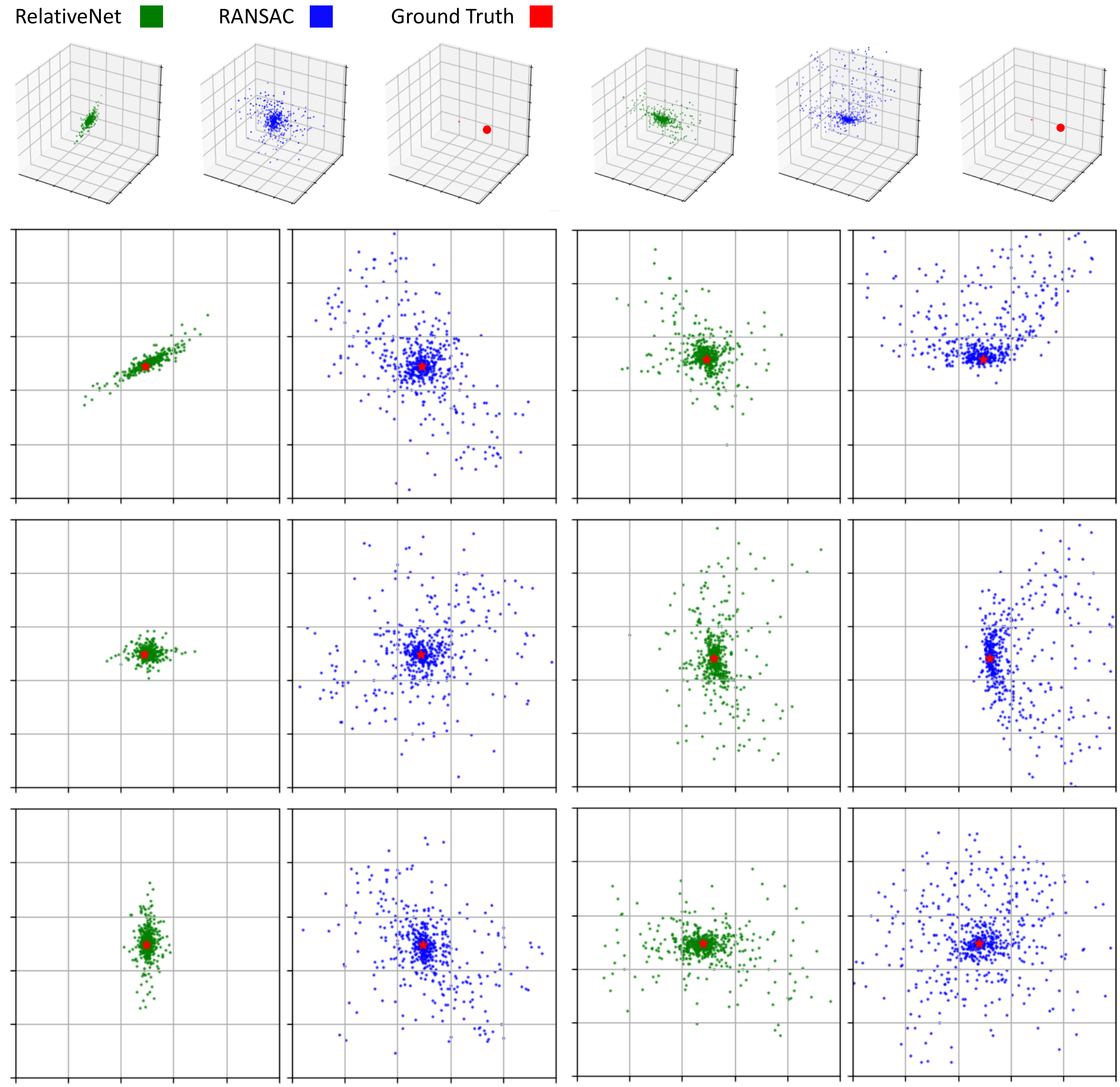
Multi-scan registration Finally, we apply our method in registering multiple scans to a common reference frame. To do that, we first align pairwise scans and obtain the most likely relative pose per pair. These poses are then fed into a global registration pipeline [2]. Note that while this method can use a global iterative closest point alignment [1] in the final stage, we deliberately omit this step to emphasize the quality of our pairwise estimates. Hence, the outcome is a rough, but nevertheless an acceptable alignment on which we can optionally apply the global-ICP refining the points and scans. The results are shown in Fig. 5 on the *Red Kitchen* sequence of the 7-scenes [4] as well as in Fig. 6 on the Sun3D Hotel sequence [5], a part of 3DMatch benchmark [6].



(a) Loosely Equivariant Feature Matching

(b) Invariant Feature Matching

Figure 2. Inlier ratio distribution of fragment pair matching result using different local features from our framework. **(a)** Matching results using equivariant features extracted by PC-FoldNet. **(b)** Matching results using invariant features extracted by PPF-FoldNet. Blue part stands for the portion of fragment pairs with correspondence inlier ratio smaller than 5%. Matching results by invariant features demonstrate a better quality for further registration procedure.



(a) Rotation Part

(b) Translation Part

Figure 3. Hypotheses distribution comparison between ones generated by RANSAC using randomly selected subset of correspondences and ones predicted by our RelativeNet. Rotation and translation parts are shown separately. The first row plots the distributions in 3D space and the following three rows are correspondent 2D projections from three different orthogonal view directions.

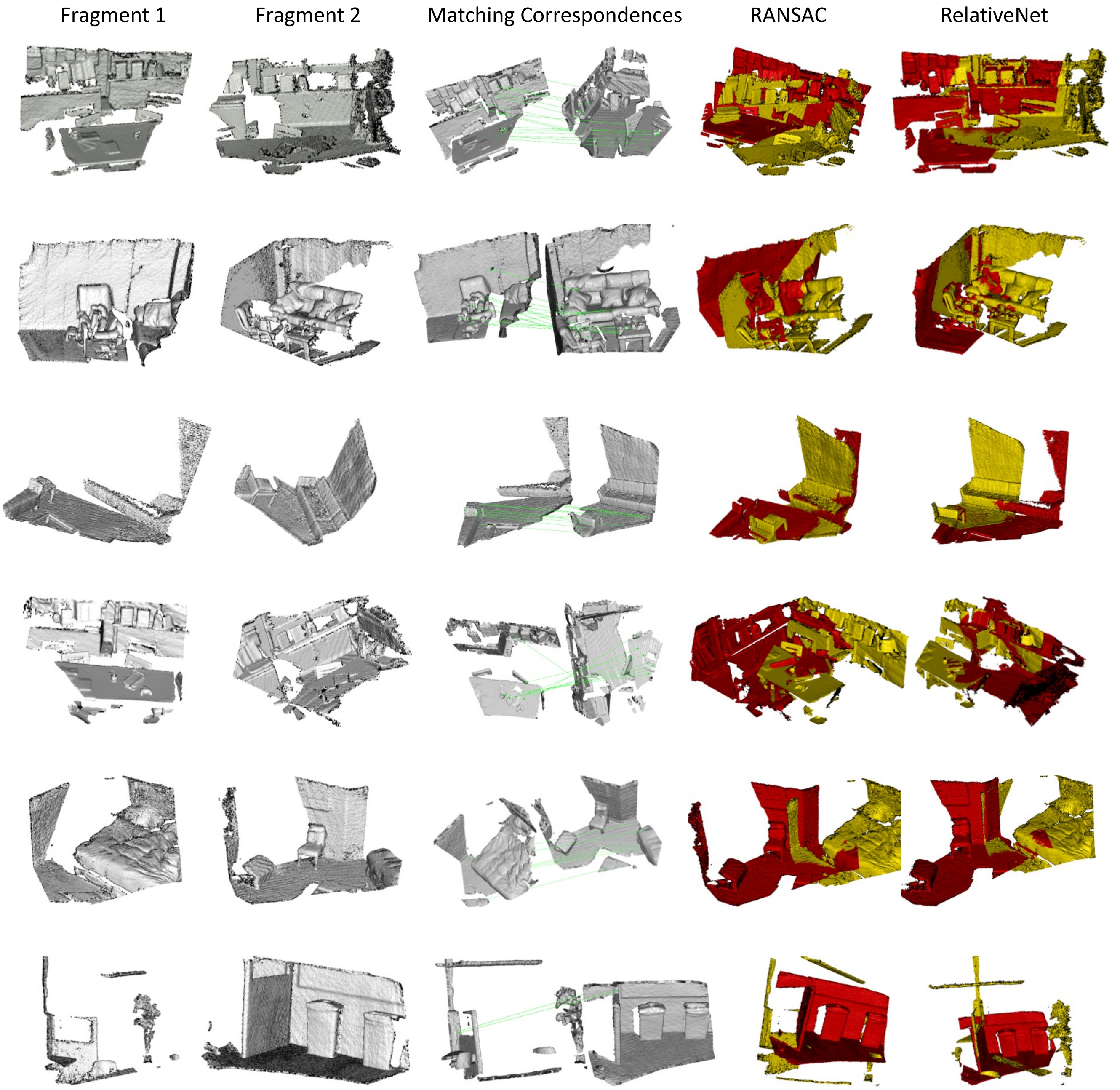
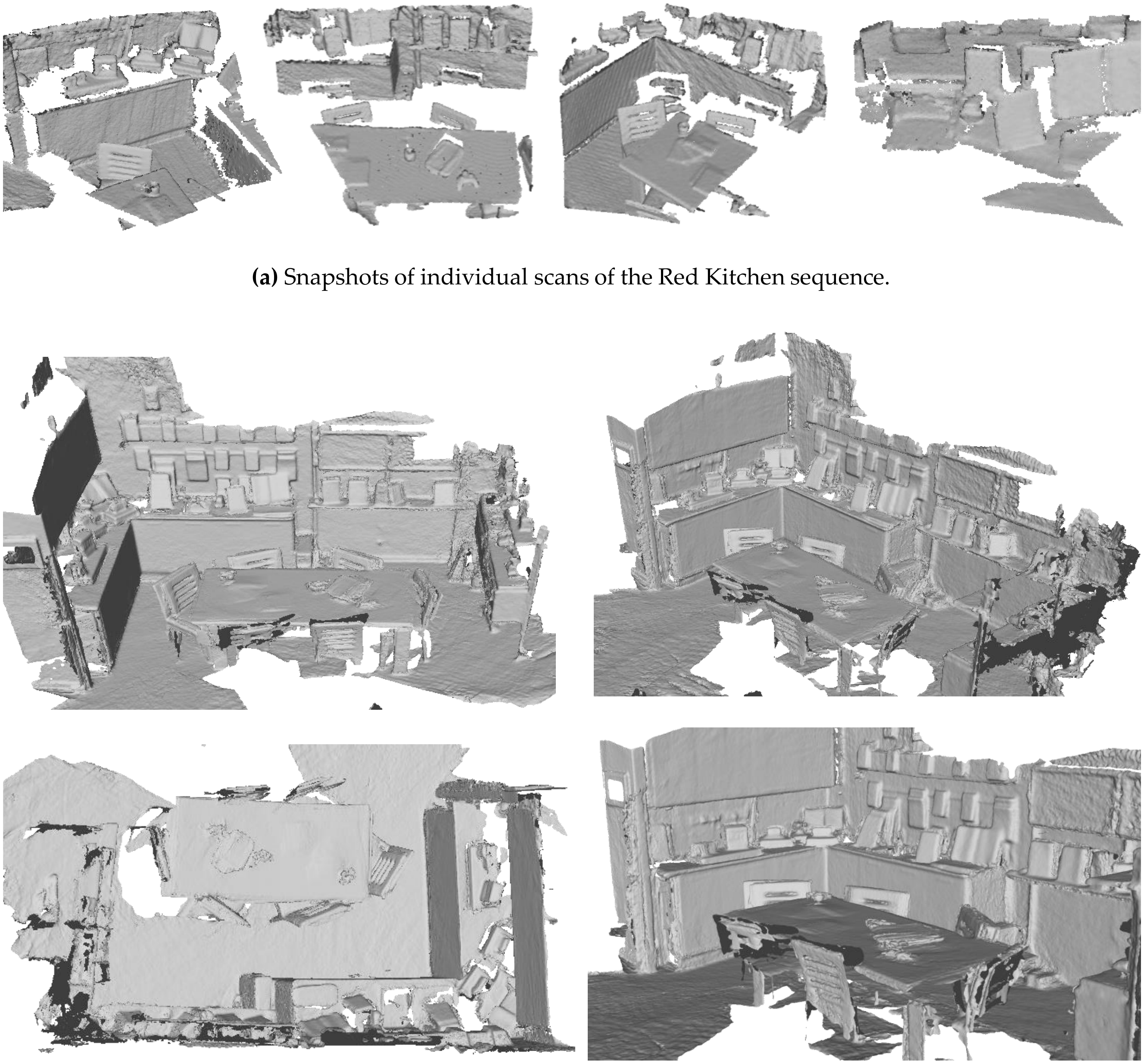


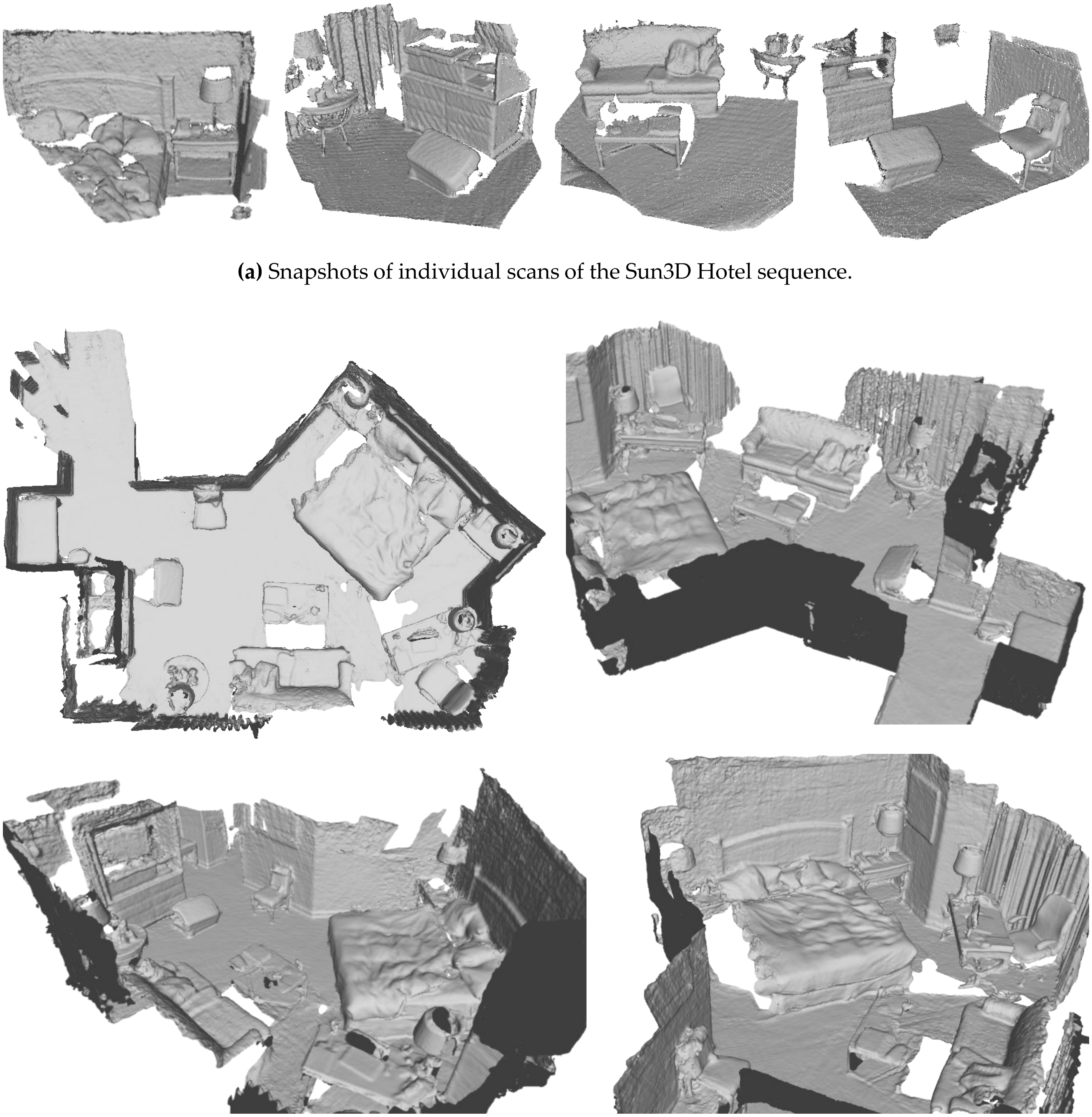
Figure 4. Some challenging fragment pairs with only a small number of correct correspondences. RANSAC fails to estimate the correct relative poses between them while our network is able to produce successful registration results. Especially, for the fragment pair in the last row, only two correct local correspondences are found, which doesn't satisfy the minimum number of inliers required by RANSAC, but still correctly handled by our method.



(a) Snapshots of individual scans of the Red Kitchen sequence.

(b) Views of the reconstruction obtained by running our method on multiple pairwise scans (No ICP)

Figure 5. Reconstruction by 3D alignment on the entire Red Kitchen sequence of the 7scenes dataset [4]. We first compute the pairwise estimates by our method and feed them into the pipeline of [2] for obtaining the poses in a globally coherent frame. Note that this dataset is a real one, acquired by a Kinect scanner. We make no assumptions on the order of acquisition.



(a) Snapshots of individual scans of the Sun3D Hotel sequence.

(b) Views of the reconstruction obtained by running our method on multiple pairwise scans (No ICP)

Figure 6. Reconstruction by 3D alignment on the entire Sun3D Hotel sequence. The reconstruction procedure is identical to the one of Fig. 5.

References

[1] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[2] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[3] H. Deng, T. Birdal, and S. Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *The European Conference on Computer Vision (ECCV)*, 2018.

[4] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013.

[5] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1625–1632, 2013.

[6] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017.