Supplementary Material for Objects Can Move: 3D Change Detection by Geometric Transformation Consistency

Aikaterini Adam[1,2], Torsten Sattler[1], Konstantinos Karantzalos[2], Tomas Pajdla[1]

[1] Czech Institute of Informatics, Robotics and Cybernetics, CTU in Prague
{Aikaterini.Adam,Torsten.Sattler,pajdla}@cvut.cz
[2] National Technical University of Athens, Greece
karank@scentral.ntua.gr

This is the supplementary material for our paper "Objects can move: 3D Change Detection by Geometric Transformation Consistency". The source code of the paper can be found at at https://github.com/katadam/ObjectsCanMove. In Section 1, we discuss more quantitative results for the task of 3D object discovery, and provide a more thorough investigation of the calculated transformations. In Section 2, more visual results of our method and the baselines are shown. We showcase corner cases where our output is inconsistent with the dataset's ground-truth annotation. We discuss in Section 3 why 3RScan [16] is the most appropriate dataset for evaluating 3D change detection and 3D object discovery, and we also evaluate our results on the sub-task of discovering added objects (Section 4).

1 Quantitative Results

Accuracy of the Computed Transformations. As explained in the main paper, the computation of the transformations induced by moving objects constitutes an essential component of the proposed method. We extract DGCNN features [17] and establish correspondences on the whole scene based on nearest neighbor search. Transformations are then computed using an iterative RANSAC procedure [2]. We further evaluate the accuracy of the rigid transformations with respect to the ground-truth ones. Results shown in Table 1, are evaluated in terms of recall, capturing the percentage of correctly calculated transformations. We also provide the Mean Translation (MTE) and Mean Rotation (MRE) Error, for all the correctly retrieved 3D rigid transforms. As in [16], we consider a transformation as successfully calculated, if the alignment errors for the translation t_Δ and rotation R_Δ are lower than $10cm, 10^\circ$ and $20cm, 20^\circ$ respectively.

We compare our approach with handcrafted and learned descriptors for establishing correspondences. The baseline methods follow the object instance relocalization protocol of the 3RScan dataset [16]: having access to an instance segmentation of the reference scan, they only need to find correspondences for 3D parts belonging to each instance. In contrast, our method does not use this supervisory signal and performs full matching between the scenes. Thus, the

Table 1. Transformation evaluation via percentage of poses within given error bounds on the position and orientation error (Recall), the Median Translation Error (MTE) (in meters), and the Median Rotation Error (MRE) (in degrees). MRE and MTE are provided for the correctly retrieved transformations, i.e., when the alignment errors for the translation t_Δ and rotation R_Δ are lower than $10cm, 10^\circ$ and $20cm, 20^\circ$ respectively

Method	Recall (<0.10m, 10°)	MRE(deg)	MTE(m)	Recall (<0.20m, 20°)	MRE(deg)	MTE(m)
FPFH [13]	2.61	7.25	0.0645	8.36	10.57	0.0776
SHOT [15]	6.79	5.35	0.0268	12.27	8.18	0.0393
3D-match dynamic [18]	5.48	5.81	0.0542	13.05	7.30	0.0708
RIO static [16]	9.92	4.33	0.0425	17.75	6.39	0.0545
RIO dynamic [16]	15.14	4.75	0.0437	23.76	6.08	0.0547
Our method	3.58	3.00	0.0799	18.21	4.25	0.1381

included baselines have access to more information compared to our approach. As expected, the baselines estimate more precise transformations. However, our results show that our approach is still competitive with such strong baselines.

Table 2. Mean IoU and mean recall for the proposed method using different number of computed transformations k to propagate geometrical consistency

Number of trans.	IoU(%)	Recall(%)
5	68.40%	76.05%
10	64.89%	77.43%
15	65.50%	79.09%

Ablation of the k Transformations Used in the Optimization. One of the tunable hyperparameters of our method is the number of top k transformations $\mathbf{T}$ used to propagate changes during the optimization. Indeed, sometimes wrong transformations with few inliers that are caused by imperfect correspondences are established. We thus ablate the top k transformations with the most inliers to propagate the change to neighboring regions. Table 2 shows the results for the proposed method when $k = 5, 10, 15$ transformations are used. As expected, using a larger k increases the mean recall (i.e., the percentage of correctly retrieved objects) and decreases the mean IoU, as in some cases, the change leads to oversegmentation. However, there is no substantial difference in the overall performance of the proposed method correlated with k. This validates the robustness of our approach.

Accuracy and Completeness of the 3D Discovered Objects. Towards assessing 3D object discovery, we also deploy the metrics of accuracy and completeness (on the point level). Per object accuracy refers to the percentage of correctly predicted 3D points out of all the points forming our discovered ob-

Table 3. Components of each presented method

Method	Init. Detect.	Comp. Transf.	Optim.
Palazzolo et al. [12] /Ours bf optim.	✓		
Taneja et al. [14]	✓		✓
Our method	✓	✓	✓

Table 4. Mean accuracy and mean completeness for the proposed method and the published baselines

Method	Acc.(%)	Compl.(%)
Palazzolo et al. [12] /Ours bf optim.	67.76 %	33.39%
Taneja et al. [14]	65.97%	30.53%
Our method	54.60%	59.20%

Table 5. Mean accuracy and mean completeness for the proposed method and the ablation study's baselines

Method	Acc.(%)	Compl.(%)
Our method	54.60%	59.20%
Ground-truth transforms.	48.79%	74.54%
Mask-RCNN	37.43%	71.39%
RANSAC inliers	52.47%	14.73%

ject. On the other hand, completeness captures how many of the ground-truth object's points were correctly retrieved by our solution. Tables 4 and 5 show the mean accuracy and mean completeness for all objects. After close examination, it is clear that our proposed method balances the most between the two metrics when compared with the two published baselines (Palazzolo et al., Taneja et al.). Table 3 summarizes the different components of each published baselines. Concerning the ablation baseline, having access to the ground-truth transformations would slightly improve the overall performance. On the other hand, assigning semantics masks to many parts of the scene (Mask-RCNN), i.e., labeling most of the scene as foreground objects, results in a high completeness rate. However, it also leads to poor accuracy as change is wrongly propagated to static regions with the same label, starting from wrong initial change detections.

2 Qualitative Results

In the following, initial detections of changing regions are depicted along with the graph cut optimization results and ground-truth annotations. Corner cases are also discussed.

Initial Detection Results. Results of the initially discovered changing regions are depicted in Figures 1, 2, and 3. It is important to note here that in most cases, the rescans constitute only partial observations of the reference scans; thus, the marked changing regions refer only to the parts visible in the rescan. After close inspection of Figures 1, 2, and 3, it becomes clear that in most cases, our initial

detection stage efficiently retrieves the changing regions. Wrong detections are primarily attributed to slight misalignments between the reference scan and the rescan.

Results after graph cut Optimization. After graph cut optimization, a 3D connected component analysis is applied both to the ground-truth annotations and to the results of our approach. This step aims to turn our detected changing regions into the final discovered objects. Figures 4, 5, and 6 present the final output of the proposed method.

Corner Cases. We also present cases where our results conflict with the annotations provided by the dataset. This could be partially attributed to unrecorded changes in the ground-truth. 3RScan [16] is a dataset built towards assessing object instance re-localization and thus does not exhaustively record every change. Typical examples are shown in Figures 6 and 9, where the non-rigid change of the curtain is not recorded. However, there are also cases where rigid transformations are not included in the ground-truth annotations (cf. the refrigerator in Figure 9). In all the above-mentioned cases, the proposed method successfully detected these changes.

The main limitation of our method is handling misalignments between scans. Indeed, differences in voxel occupancy and depth values occur when the reference scan and the rescan are not correctly registered. Thus, regions are falsely labeled as changing during our method's initial detection step. A typical example is Figure 8, where parts of the floor are wrongly retrieved as changing due to the misalignment. Post-processing steps such as matching against a 3D object database, can eliminate these false detections.

Baseline Comparisons. As stated in the Section 4 of the main paper, we compare our method against two published baselines, Palazzolo et al. [12] and Taneja et al. [14]. We also compare against three baselines in the form of an ablation study (Ground-truth transforms, Mask-RCNN, RANSAC Inliers). Qualitative results for the two published baselines and the ablation baselines (Ground-truth transforms and Mask-RCNN) are depicted in Figure 7.

The main paper explains that Palazzolo et al. [12] is equivalent to our method before the optimization step, i.e., it predicts change through depth comparison. The depicted visual results prove the need for a more sophisticated solution rather than simply relying on inconsistencies between 2D projections. On the other hand, Taneja et al. [14] performs a graph cut optimization, where the binary term ensures photoconsistency. This constraint seems to perform well on objects with homogeneous texture, such as the chair in Figure 7, but fails when the same object has multiple textures (cf. the cabinet in Figure 7).

Moving on to the ablation study, the baseline using the ground-truth transformations provided by the dataset seems to work better than the presented method. This is expected since our method firstly computes transformations and then discovers objects. Thus, for the proposed method, the uncertainty of the calculated transformations is propagated to the results of the graph optimization step. Finally, it is evident that when Mask-RCNN is used, a lot of

Table 6. Applicable datasets for change detection/ object discovery

Method	Scans	Rescans	Instance Seg.	Annotation	Available
Finman et al.[4]	2	67	?	?	—
Langer et al.[10]	1	4	—	—	—
Katsura et al.[8]	2	10+?	?	?	—
Herbst et al.[7]	4	24	✓	✓	—[3]
Mason et al.[11]	1	67	—	—	✓[4]
Ambrus et al.[1]	1	88	—	—[5]	✓
Fehr et al.[3]	3	23	—	—	✓
Wald et al.[16] - 3RScan	478	1482	✓	✓	✓[6]
Halber et al.[6]	13	45	✓	—	✓
Langer et al.[9]	5	31	—	✓	✓

background regions are labelled as changing, since they were wrongly extracted as foreground in the RGB-D images (i.e., they were incorrectly assigned a semantic mask). Moreover, it completely fails in regions that were not detected as a foreground object (cf. cabinet in Figure 7), due to the lack of relevant training data for these concepts. In contrast, our method does not rely on any predefined notion of what an object should look like.

3 Datasets

The 3RScan dataset [16] is a dataset built towards assessing object instance relocalization under rigid transformations. Thus, the dataset provides information about the transformations induced by moving objects, along with an instance segmentation of each scene. It also provides information about non-rigid changes in some instances. Hence, we generate the ground-truth annotations by (i) extracting all the information about rigid and non-rigid movements as provided by the dataset, and (ii) by comparing the instance segmentation of the reference scan and rescan to discover objects that have been added or removed.

As stated above, even though 3RScan is not directly designed for 3D change detection/ 3D object discovery, ground-truth information can be generated without manual labeling. It is also a very large and diverse dataset, and thus, to the best of our knowledge, the most appropriate dataset for our needs. Table 6 summarizes the characteristics of the adopted dataset, against other applicable datasets that were not suitable for our scenario.

As shown in Table 6, there is no widely available dataset appropriate for evaluating 3D indoor change detection and 3D object discovery. Each work evaluates its efficiency on tailor-made datasets, some of which [4,10,8,7] are not

[3] The provided URL is not valid
[4] Data available upon request
[5] Inconsistent and incomplete annotation
[6] With our provided code

publicly available. Concerning the public datasets, [11] provides a relative large amount of rescans. However, only a single environment is considered. Moreover, its objects are not annotated so it cannot be used for quantitative evaluation. Similarly, [1] consists of data captured by a robot in an office setting. The diversity of the provided scenes is thus limited. The annotation is mostly inconsistent, since some selected objects are annotated as new while other objects that are physically new in a scene are not. In a similar vein to our used dataset [16], [3] uses a hand-held Google Tango device to capture three rooms (reference scans) and 23 rescans. Taking into account that the dataset is much smaller and less diverse and its complete lack of annotations, we have decided not to use it. The authors of [9] have created their own dataset, for evaluating added small objects (from the YCB dataset [5]) in the scene. [10] provides ground-truth annotation only for novel objects and not for moved ones, which makes the quantitative evaluation of our task hard as not all the cases we are interested in can be directly tested. Finally, [6] aims at tracking instance segmentation across temporal changes. Thus, a ground-truth instance annotation is provided for every rescan, but no annotations concerning moving/static objects.

4 Novel Objects

Novel objects are also of broad interest to multiple robotic applications. Since existing works in the research community [10,9] focus on novel added objects, we have decided to create a subtask of discovering all added objects in the scene. To this end, we have prepossessed 3RScan to create a new ground truth including only the novel objects, and we also decided to compare our algorithm against [10]. Originally, [10] detects only small objects on the floor. We modified this work to discover objects regardless of their size and position for a fair comparison. Table 7 shows the results in terms of IoU and recall at the voxel level, to capture the volumetric overlap between ground truth and predictions. After close inspection of Tab. 7, it is clear that our method outperforms the most competitive baseline in the sub-task of discovering added objects.

Table 7. Metrics of the proposed method and [10] on novel objects of the 3RScan dataset

Method	IoU(%)	Recall(%)
[10]	73.85%	64.25%
Our method	75.08%	76.70%

References

1. Ambrus, R., Folkesson, J., Jensfelt, P.: Unsupervised object segmentation through change detection in a long term autonomy scenario. In: 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). pp. 1181–1187. IEEE (2016)

2. Bolles, R.C., Fischler, M.A.: A ransac-based approach to model fitting and its application to finding cylinders in range data. In: IJCAI. vol. 1981, pp. 637–643. Citeseer (1981)
3. Fehr, M., Furrer, F., Dryanovski, I., Sturm, J., Gilitschenski, I., Siegwart, R., Cadena, C.: Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In: 2017 IEEE International Conference on Robotics and automation (ICRA). pp. 5237–5244. IEEE (2017)
4. Finman, R., Whelan, T., Kaess, M., Leonard, J.J.: Toward lifelong object segmentation from change detection in dense rgb-d maps. In: 2013 European Conference on Mobile Robots. pp. 178–185. IEEE (2013)
5. Gadde, R., Jampani, V., Marlet, R., Gehler, P.V.: Efficient 2d and 3d facade segmentation using auto-context. IEEE transactions on pattern analysis and machine intelligence **40**(5), 1273–1280 (2017)
6. Halber, M., Shi, Y., Xu, K., Funkhouser, T.: Rescan: Inductive instance segmentation for indoor rgbd scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2541–2550 (2019)
7. Herbst, E., Henry, P., Ren, X., Fox, D.: Toward object discovery and modeling via 3-d scene comparison. In: 2011 IEEE International Conference on Robotics and Automation. pp. 2623–2629. IEEE (2011)
8. Katsura, U., Matsumoto, K., Kawamura, A., Ishigami, T., Okada, T., Kurazume, R.: Spatial change detection using normal distributions transform. ROBOMECH Journal **6**(1), 1–13 (2019)
9. Langer, E., Patten, T., Vincze, M.: Robust and efficient object change detection by combining global semantic information and local geometric verification. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8453–8460. IEEE (2020)
10. Langer, E., Ridder, B., Cashmore, M., Magazzeni, D., Zillich, M., Vincze, M.: On-the-fly detection of novel objects in indoor environments. In: 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 900–907. IEEE (2017)
11. Mason, J., Marthi, B.: An object-based semantic world model for long-term change detection and semantic querying. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3851–3858. IEEE (2012)
12. Palazzolo, E., Stachniss, C.: Fast image-based geometric change detection given a 3d model. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 6308–6315. IEEE (2018)
13. Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (fpfh) for 3d registration. In: 2009 IEEE international conference on robotics and automation. pp. 3212–3217. IEEE (2009)
14. Taneja, A., Ballan, L., Pollefeys, M.: Image based detection of geometric changes in urban environments. In: 2011 International Conference on Computer Vision. pp. 2336–2343. IEEE (2011)
15. Tombari, F., Salti, S., Di Stefano, L.: Unique signatures of histograms for local surface description. In: European conference on computer vision. pp. 356–369. Springer (2010)
16. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7658–7667 (2019)
17. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog) **38**(5), 1–12 (2019)

18. 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)

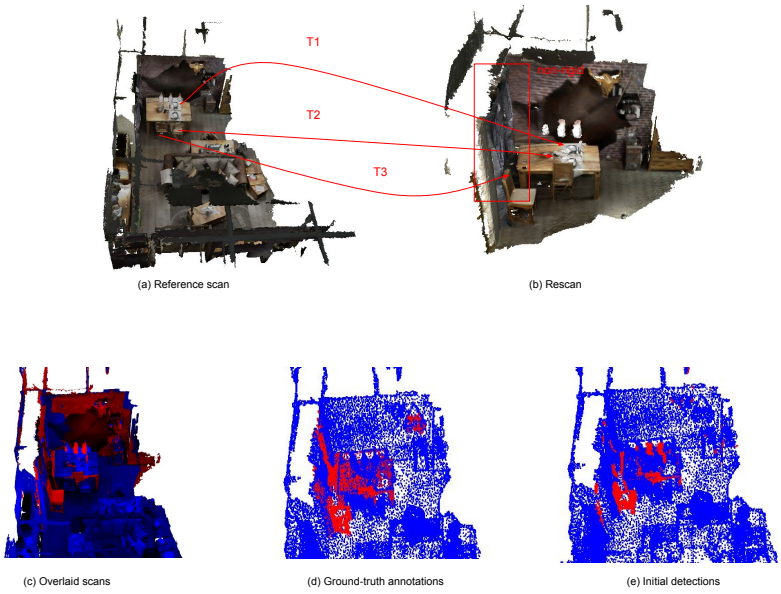

Fig. 1. Reference scans (a) and re-scans (b) of multiple scenes with highlighted occurred changes (T1, T2, etc. refer to the ground-truth transformations between moving objects). Overlaid meshes of reference scan and rescan in red and blue (c). Ground-truth changed regions of the point clouds (in red) overlaid on the reference scan (in blue) (d). Initial change detection results of the point clouds (in red) overlaid on the reference scan (in blue) (e)

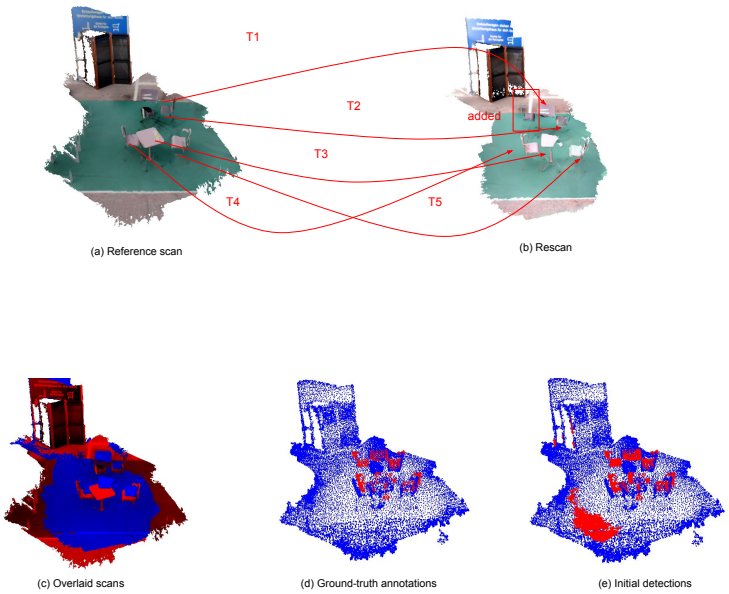

Fig. 2. Reference scans (a) and re-scans (b) of multiple scenes with highlighted occurred changes (T1, T2, etc. refer to the ground-truth transformations between moved objects). Overlaid meshes of reference scan and rescan in red and blue (c). Ground-truth changed regions of the point clouds (in red) overlaid on the reference scan (in blue) (d). Initial change detection results of the point clouds (in red) overlaid on the reference scan (in blue) (e)

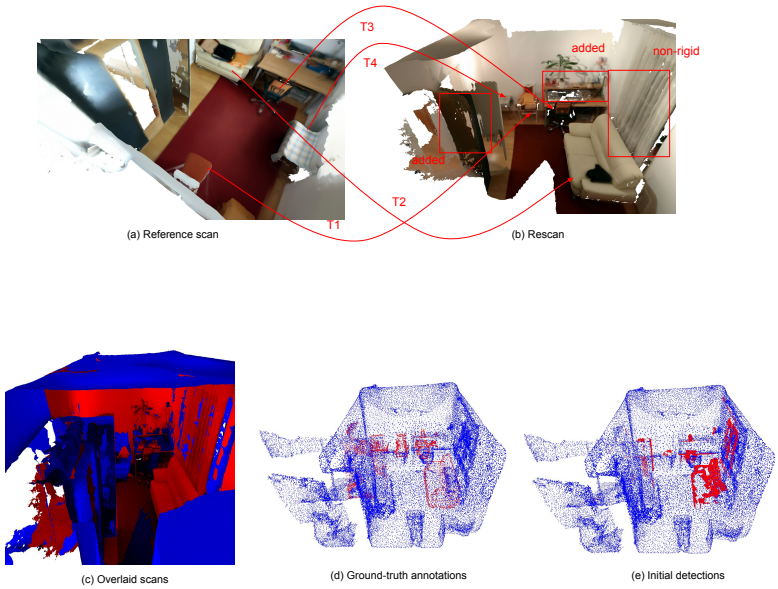

Fig. 3. Reference scans (a) and re-scans (b) of multiple scenes with highlighted occurred changes (T1, T2, etc.. refer to the ground-truth transformations between moved objects). Overlaid meshes of reference scan and rescan in red and blue (c). Ground-truth changed regions of the point clouds (in red) overlaid on the reference scan (in blue) (d). Initial change detection results of the point clouds (in red) overlaid on the reference scan (in blue) (e)

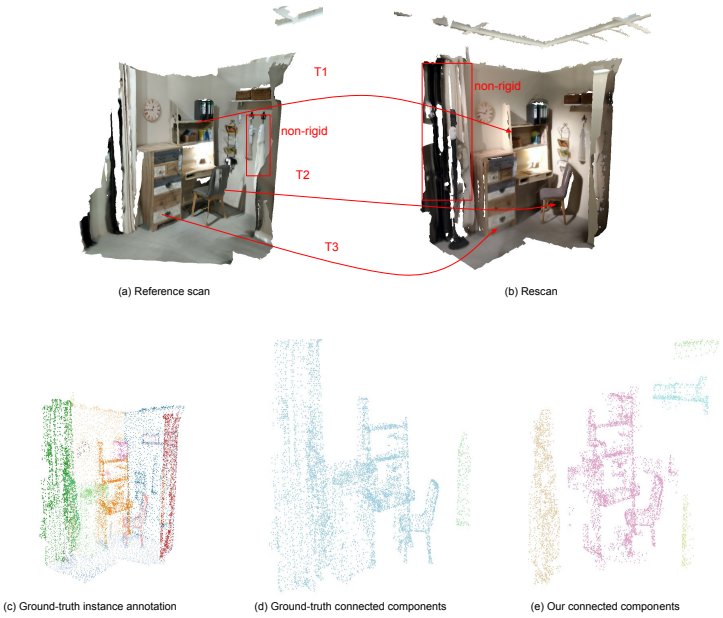

Fig. 4. The meshes of the reference scan (a) and the rescan (b). Ground-truth instance segmentation of the point cloud of the rescan (c). Ground-truth connected components of the point cloud (d), compared with the connected components of our solution in (e)

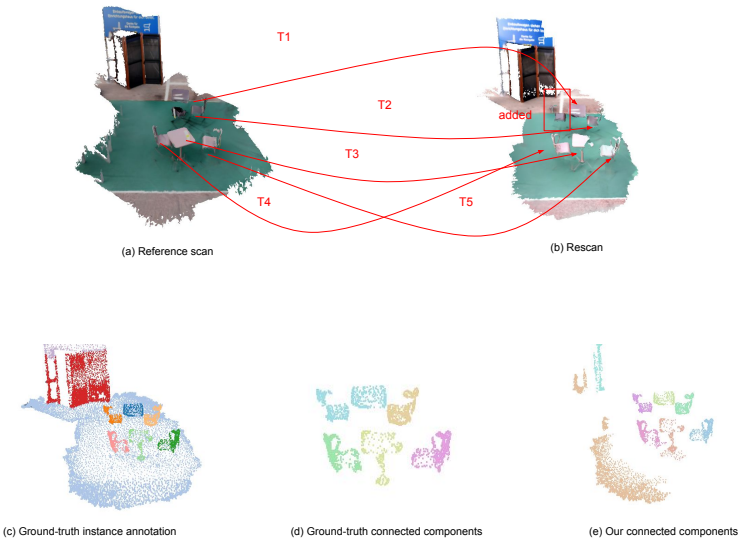

Fig. 5. The meshes of the reference scan (a) and the rescan (b). Ground-truth instance segmentation of the point cloud of the rescan (c). Ground-truth connected components of the point cloud (d), compared with the connected components of our solution in (e)

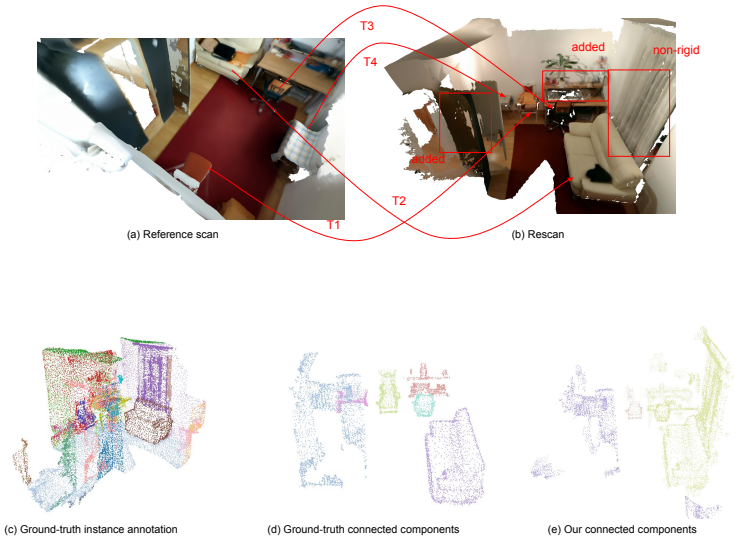

Fig. 6. The meshes of the reference scan (a) and the rescan (b). Ground-truth instance segmentation of the point cloud of the rescan (c). Ground-truth connected components of the point cloud (d), compared with the connected components of our solution in (e)

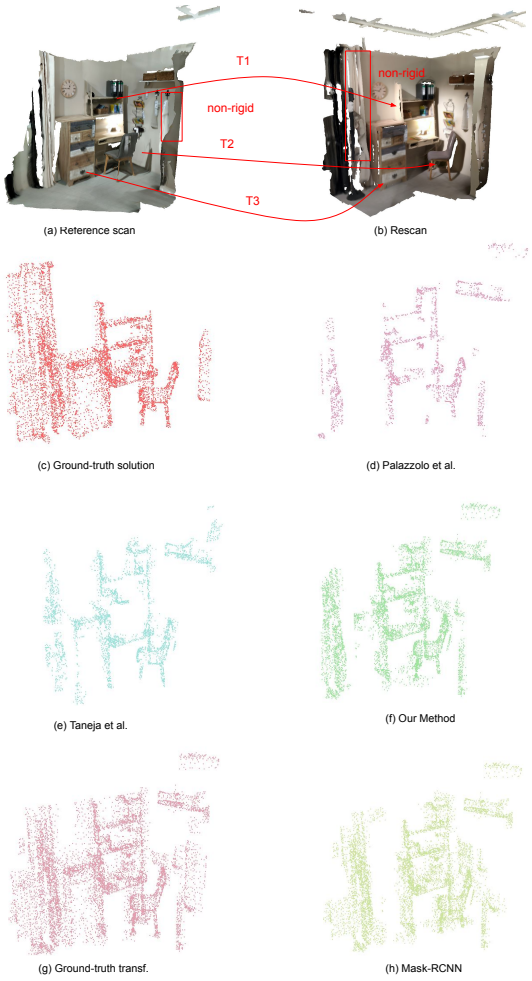

Fig. 7. The meshes of the reference scan (a) and the rescan (b). Ground-truth solution (c). Results of the published baselines in (d) and (e). Results of the proposed method in (f). Results of the ablation study's baselines in (g) and (h)

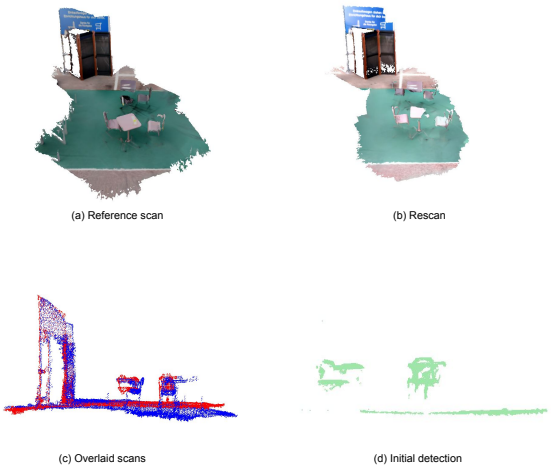

Fig. 8. Slight misalignment between reference scan (a) and rescan (b). Overlaid reference scan and rescan (depicted in blue and red respectively) in (c). Initial detection results in (d)

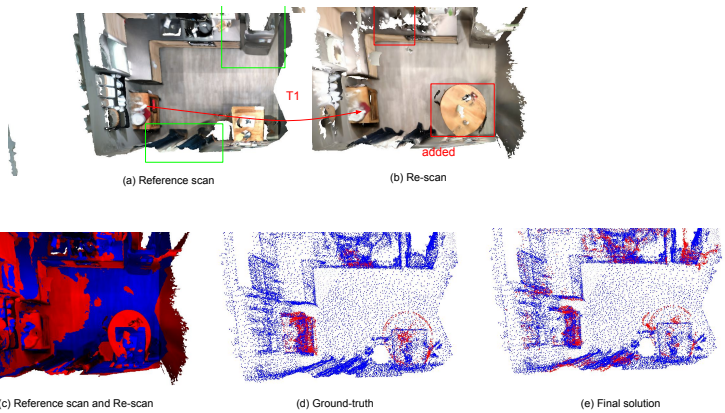

Fig. 9. Changes not recorded in the ground-truth. The mesh of the reference scan is depicted in (a), and the green bounding boxes underline unrecorded changes. The mesh of the rescan in (b). Recorded changes in the ground-truth are underlined in red color. The meshes of the two scans overlaid in red and blue color, respectively (c). Ground-truth annotations (in red) overlaid on the point cloud of the reference scan in blue (d), and our graph cut optimization results (in red) overlaid on the point cloud of the reference scan in blue (e). The rigid change of the refrigerator and the non-rigid change of the curtain that were not recorded in the ground-truth are successfully detected by our method