

RIO: 3D Object Instance Re-Localization in Changing Indoor Environments

Johanna Wald¹ Armen Avetisyan¹ Nassir Navab¹ Federico Tombari^{1,2,*} Matthias Nießner^{1,*}
¹ Technical University of Munich ² Google

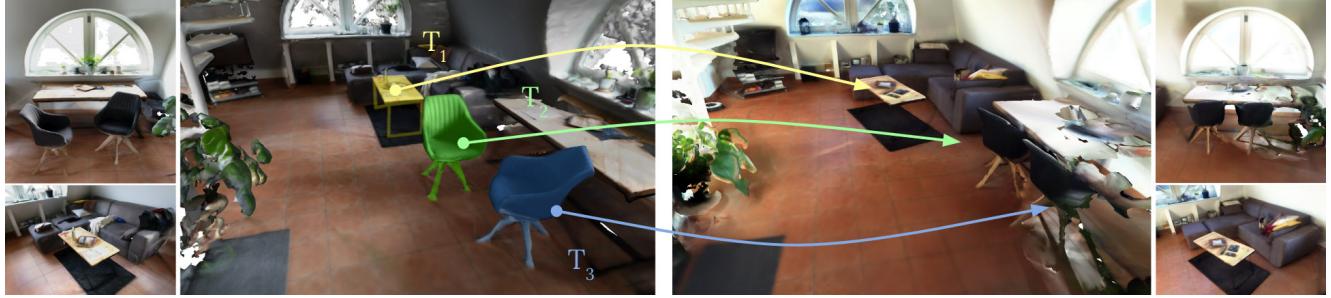


Figure 1: 3D object instance re-localization benchmark: we want to robustly estimate the 6DoF pose ($\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_n$) of changed rigid object instances from a segmented source to a target scan taken at a later point in time.

Abstract

In this work, we introduce the task of 3D object instance re-localization (RIO): given one or multiple objects in an RGB-D scan, we want to estimate their corresponding 6DoF poses in another 3D scan of the same environment taken at a later point in time. We consider RIO a particularly important task in 3D vision since it enables a wide range of practical applications, including AI-assistants or robots that are asked to find a specific object in a 3D scene. To address this problem, we first introduce 3RScan, a novel dataset and benchmark, which features 1482 RGB-D scans of 478 environments across multiple time steps. Each scene includes several objects whose positions change over time, together with ground truth annotations of object instances and their respective 6DoF mappings among re-scans. Automatically finding 6DoF object poses leads to a particular challenging feature matching task due to varying partial observations and changes in the surrounding context. To this end, we introduce a new data-driven approach that efficiently finds matching features using a fully-convolutional 3D correspondence network operating on multiple spatial scales. Combined with a 6DoF pose optimization, our method outperforms state-of-the-art baselines on our newly-established benchmark, achieving an accuracy of 30.58%.

1. Introduction

3D scanning and understanding of indoor environments is a fundamental research direction in computer vision laying the foundation for a large variety of applications ranging from indoor robotics to augmented and virtual reality. In particular, the rapid progress in RGB-D scanning systems [17, 18, 31, 6] allows to obtain 3D reconstructions of indoor scenes using only low-cost scanning devices such as the Microsoft Kinect, Intel Real Sense, or Google Tango. Along with the ability to capture 3D maps, researchers have shown significant interest in using these representations to perform 3D scene understanding and developed a rapidly-emerging line of research focusing on tasks such as 3D semantic segmentation [4, 19, 26] or 3D instance segmentation [10]. However, the shared commonality between these works is that they only consider static scene environments. In this work, we focus on environments that change over time. Specifically, we introduce the task of object instance re-localization (RIO): given one or multiple objects in an RGB-D scan, we want to estimate their corresponding 6DoF poses in another 3D scan of the same environment taken at a different point in time. Therefore, the captured reconstructions naturally cover a variety of temporal changes; see Fig. 1. We believe this is a critical task for many indoor applications, for instance, for a robot or virtual assistant to find a specific object in its surrounding environment.

The main challenge in RIO – finding the 6DoF of each object – lies in establishing good correspondences between re-scans, which is non-trivial due to different scanning pat-

* Authors share senior authorship.

<https://waldjohannau.github.io/RIO/>

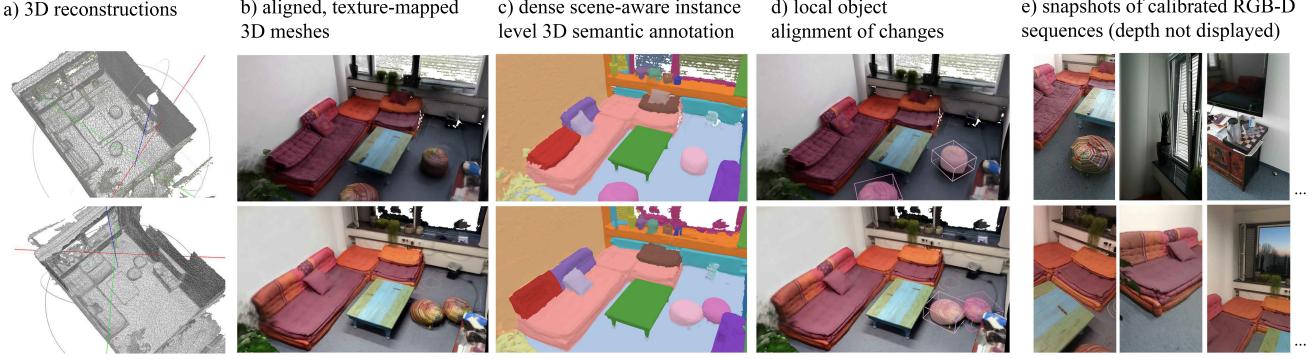


Figure 2: Example of a 3D scene pair of the 3RScan dataset. It provides: e) calibrated RGB-D sequences, a), b) aligned textured 3D reconstructions, c) dense instance-level semantic segmentation as well as d) symmetry-aware local alignment of changes.

terns and changing geometric context. These make the use of hand-crafted geometric descriptors, such as FPFH [22] or SHOT [27], less effective. Similarly, learned 3D feature matching approaches, such as 3DMatch [32, 7], cannot be easily leveraged since they are trained on self-supervised correspondences from static 3D scenes, and are hence very susceptible to geometry changes. One of the major limitations in using data-driven approaches for object instance localization is the scarce availability of supervised training data. While existing RGB-D datasets, such as ScanNet [4] or SUN RGB-D [25], provide semantic segmentations for hundreds of scenes, they lack temporal annotations across scene changes. In order to address this shortcoming, we introduce 3RScan a new dataset that is composed of 1482 RGB-D sequences. An essential novelty of the proposed dataset is that several re-scans are provided for every environment. The dataset includes not only dense ground truth semantic instance annotations (for every scan), but also associates objects that have changed in appearance and/or location between re-scans. In addition to using 3RScan for training feature descriptors, we also introduce a new benchmark for object instance localization.

In order to learn from this data, we propose a fully-convolutional multi-scale network capable of learning geometric features in dynamic environments. The network is trained with corresponding TSDF (truncated signed distance function) patches on moved objects extracted at two different spatial scales in a self-supervised fashion. As a result, we obtain change-invariant local features that outperform state-of-the-art baselines in correspondence matching and on our newly created benchmark for re-localization of object instances. In summary, we explore the task of 3D Object Instance Re-Localization in changing environments and contribute:

- 3RScan, a large indoor RGB-D dataset of changing environments that are scanned multiple times. We provide

ground truth annotations for dense semantic instance labels and changed object associations.

- a new data-driven object instance re-localization approach that learns robust features in changing 3D contexts based on a geometric multi-scale neural network.

2. Related Work

3D Object Localization and Keypoint Matching 3D object localization and pose estimation via keypoint matching are long standing areas of interest in computer vision. Until very recently, 3D hand-crafted descriptors [27, 22] were prominently used to localize objects under occlusion and clutter by determining 3D point-to-point correspondences. However with the success of machine learning, the interest shifted to deep learned 3D feature descriptors capable of embedding 3D data, such as meshes or point clouds, in a discriminative latent space [32, 19, 7]. Even though these approaches show impressive results on tasks such as correspondences matching and registration, they are restricted to static environments. In this work, we go one step further by focusing on dynamic tasks; specifically, we aim to localize given 3D objects from a source scan in a cluttered target scan which contains common geometric and appearance changes.

RGB-D Scene Understanding Scene understanding methods based on RGB-D data generally rely on volumetric or surfel-based SLAM to reconstruct the 3D geometry of the scene while fusing semantic segments extracted via Random Forests [29, 30] or CNNs [13, 15]. Other works such as SLAM++ [24] or Fusion++ [12] operate on an object level and create semantic scene graphs for SLAM and loop closure. Non-incremental scene understanding methods, in contrast, process a 3D scan directly to obtain semantic, instance or part segmentation [19, 20, 21, 5, 10]. Independently from the approach, all these methods rely on the assumption that

Table 1: RGB-D indoor datasets for 3D scene understanding: We list synthetic as well as real dataset and compare their size together with other properties such as the availability of scene changes.

Dataset	Size	Real	Data Acquisition / Generation	Benchmarks	Changes
NYUv2 [16]	464 scenes	✓	recordings with Kinect	Depth and Semantics	✗
SUN RGB-D [25]	10k frames	✓	recordings with 4 different sensors	3D Object Detection	✗
SUN-CG [26]	45K rooms, 500K images	✗	rendered, layout hand-designed	Scene-Completion	✗
ScanNet [4]	1513 scans, 2.5M images	✓	recordings with Structure Sensor	Semantic Voxel Labeling	✗
Fehr et al. [8]	23 scans of 3 scenes	✓	recordings with Tango	Change Detection	✓
Matterport3D [3]	90 buildings, ~ 200k images	✓	recordings with Matterport	several	✗
SceneNet RGB-D [14]	15K trajectories, 5M images	✗	photo-realistic, random scenes	SLAM	✗
InteriorNet [11]	<i>millions</i> / unknown	✗	photo-realistic, layout hand-designed	SLAM	✓
RGB Reloc [28]	4 scenes, 12 rooms	✓	recording with Kinect	Camera Re-Localization	✗
3RScan (Ours)	1482 scans of 478 scenes	✓	recordings with Tango	Object Instance Re-Localization	✓

objects are static and the scene structure does not change over time.

RGB-D Datasets Driven by the great interest in the development of scene understanding applications, several large-scale datasets based on RGB-D data have been recently proposed [9]. We have summarized the most prominent efforts in Table 1, together with their main features (*e.g.*, number of scenes, mean of acquisition). The majority of datasets do not include changes in the scene layout and objects therein, and assume each scene is static over time. This is the case of ScanNet [4], currently the largest real dataset for indoor scene understanding consisting of 1500 scans of approx. 750 unique scenes. Notably, only a few recent proposals started exploring the idea of collecting scene changes to allow long-term scene understanding. InteriorNet [11] is a large-scale synthetic dataset, in which random physics-based furniture shuffles and illumination changes are applied to generate appearance and geometry variations which indoor scenes typically undergo. Several state-of-the-art sparse and dense SLAM approaches are compared on this benchmark. Despite the impressive size and indisputable usefulness, we argue that, due to the domain gap between real and synthetic imagery, the availability of real sequences remain crucial for the development of long-term scene understanding. To the best of our knowledge, the only real dataset encompassing scene changes is the one released by Fehr *et al.* [8], which includes 23 sequences of 3 different rooms used to segment the scene structure from the movable furniture, though lacking the annotations and necessary size to train and test current learned approaches.

3. 3RScan-Dataset

We propose **3RScan**, a large scale, Real-world dataset which contains multiple (2 – 12) 3D snapshots (**Re-scans**) of naturally changing indoor environments, designed for benchmarking emerging tasks such as long-term SLAM,

scene change detection [8] and camera or object instance Re-Localization. In this section, we describe the acquisition of the scene scans under dynamic layout and moving objects, as well as that of annotation in terms of object pose and semantic segmentation.

3.1. Overview

The recorded sequences are either a) controlled, where pairs are acquired within a time frame of only a few minutes under known scene changes or b) uncontrolled, where unknown changes naturally occurred over time (up to a few months) via scene-user interaction. All 1482 sequences were recorded with a Tango mobile application to enable easy usage for untrained users. Each sequence was processed offline to get bundle-adjusted camera poses with loop-closure and texture mapped 3D reconstructions. To ensure high variability, 45+ different people recorded data in more than 13 different countries. Each sequence comes with aligned semantically annotated 3D data and corresponding 2D frames (approximately 363k in total), containing in detail:

- calibrated RGB-D sequences with variable n RGB $\mathbf{R}_i, \dots, \mathbf{R}_n$ and depth images $\mathbf{D}_i, \dots, \mathbf{D}_n$.
- textured 3D meshes
- camera poses $\mathbf{P}_i, \dots, \mathbf{P}_n$ and calibration parameters \mathbf{K} .
- global alignment among scans from the same scene as a global transformation \mathbf{T} .
- dense instance-level semantic segmentation where each instance has a fixed ID that is kept consistent across different sequences of the same environment.
- object alignment, *i.e.* a ground truth transformation $\mathbf{T}_{GT} = \mathbf{R}_{GT} + \mathbf{t}_{GT}$ for each changed object together with its symmetry property.

- intra-class transformations A of ambiguous instances in the reference to recover all valid object poses in the re-scans (see Figure 3).

3.2. Scene Changes

Due to the repetitive recording of interactive indoor environments, our data naturally captures a large variety of temporal scene changes. Those changes are mostly rigid and include a) objects being moved (from a few centimeters up to a few meters) or b) objects being removed or added to the scene. Additionally, non-rigid objects such as curtains or blankets and the presence of lighting changes create additional challenging scenarios.

3.3. Annotation

The dataset comes with rich annotations which include scan-to-scene-mappings and 3D transformations (section 3.3.2) together with dense instance segmentation (section 3.3.1). More details and statistics regarding the annotations are given in the supplementary material.

3.3.1 Semantic Segmentation

Similarly to ScanNet [4], instance-level semantic annotations are obtained by labeling on a segmented 3D surface directly. For this, each reference scan was annotated with a modified version of ScanNet’s publicly available annotation framework. To reduce annotation time, we propagate the annotations in a segment-based fashion from the reference scan to each re-scan using the global alignment T with the scan-to-scene mappings. This gives us very good annotation estimates for the re-scans, with the assumption that most parts of the scene remain static. Figure 4 gives an example of automatic label propagation from a hand-annotated scene in the presence of noise and scene changes. Semantic segments were annotated by human experts using a web-based crowd-sourcing interface and verified by the authors. The average annotation coverage of the semantic segmentation for the entire dataset is 98.5%.

3.3.2 Instance Changes

To obtain instance-level 3D transformations, a keypoint-based 3D annotation and verification interface was developed based on the CAD alignment tool used in [2]. A 3D transformation is obtained by applying Procrustes on manually annotated 3D keypoint correspondences on the object from the reference and its counterpart in the re-scan (see Figure 5). Additionally to this 3D transformation, a symmetry property was assigned to each instance.

3.4. Benchmark

Based on this data, we set up a public benchmark for 3D instance-level object re-localization in changing indoor environments. Given one or multiple objects in a segmented source scene, we want to estimate the corresponding 6DoF poses in a target scan of the same environment taken at a different point in time. Namely, transformations $T_1 = R_1 + t_1, \dots, T_m$ as a translation t_1, \dots, t_m and rotation R_1, \dots, R_m need to be detected for all given m instances in A (left Figure 1) to instances in B (right). Predictions are evaluated against the annotated 3D transformation. A 6DoF pose estimation is considered successful if the translation and rotation error to the given ground truth transformation is within a small range. In our experiments we set these thresholds to $t \leq 10\text{ cm}$ and $r \leq 10^\circ$ and $t \leq 20\text{ cm}$ and $r \leq 20^\circ$ respectively. To avoid misalignment of symmetric objects, the respective symmetry property is considered. We publicly release our dataset with a standardized test, validation and training set (see Table 2) and all mentioned annotations. To allow a fair comparison of different methods, we also release a hidden test set together with an automatic server-side testing script.

Table 2: Statistics on the test, train and validation set of 3RScan.

	test	train	validation	total
#scenes	46	385	47	478
#re-scans	101	793	110	1004
#scans	147	1178	157	1482

4. 3D Object Instance Re-Localization

In order to address the task of RIO, we propose a new data-driven approach that finds matching features in changing 3D scans using a 3D correspondence network. Our network operates on multiple spatial scales to encode change-invariant neighborhood information around the object and the scene. Object instances are re-localized by combining the learned correspondences with RANSAC and a 6DoF object pose optimization.

4.1. Data Representation

The input of our network are TSDF patches. For each 3D keypoint on the source object or the target scene, the surrounding 3D volume is extracted at two different scales. They are chosen to be $32 \times 32 \times 32$ voxel grids containing TSDF values with spatial resolutions of $(1.2\text{ m})^3$ and $(0.6\text{ m})^3$. Their corresponding voxel sizes are 1.875 cm and 3.75 cm.

4.2. Network Architecture

The network architecture of RIO is visualized in Figure 6. Due to non-padded convolutions and two pooling layers the



Figure 3: Instance ambiguities in presence of scene changes: since the instance mapping is unknown, multiple solutions are plausible, which we provide in our dataset from user annotations indicating all possibilities.

input volumes are reduced to a 512-dimensional feature vector. It consists of two separate single scale encoders (SSE) and a subsequent multi-scale encoder (MSE). The two different input resolutions capture different neighborhoods with a different level of detail. Since both single scale encoder branches are identical, their network responses are concatenated before being fed into the MSE, as visualized in Figure 6. This multi-scale architecture helps to simultaneously capture fine geometric details as well as higher-level semantics of the surroundings. We show that our multi-resolution network produces richer features and therefore outperforms single scale architectures that process each scale independently by a large margin. Please also note that the two network branches do not share weights since they pro-



Figure 4: Propagation result from a reference (left) to a re-scan (center) and the manual cleanup (right). Please note the false propagations in the presence of scenes changes: here the orange armchair was moved and its label was therefore incorrectly propagated.



Figure 5: Example of the correspondences-based 3D instance alignment (right). 3D transformations are computed by manual annotation of corresponding keypoints on the objects (left, green) and the scene (center, red) respectively.

cess the geometry of different context. To achieve a strong gradient near the object surface the raw TSDF is inverted in the first layer of the network such that

$$\hat{TSDF} = 1 - |TSDF|. \quad (1)$$

4.3. Training

During training, a triplet network architecture together with a triplet loss (equation 2) is used. It maximizes the L_2 distance of negative patches and minimizes the L_2 distance of positive patches. We choose the margin α to be 1. For optimization, Adam optimizer with an initial learning rate of 0.001 is used.

$$\sum_{i=1}^N N \left[\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha \right] \quad (2)$$

4.4. Training Data: From Static to Dynamic

We initially train our network fully self-supervised with static TSDF patches extracted from RGB-D sequences of our dataset. To be able to deal with partial reconstructions induced by different scanning patterns, two small sets of non-overlapping frames are processed to produce two different TSDF volumes of the same scene. Then, first Harris 3D keypoints are extracted on one volume, then these same locations are refined on the other volume via non-maxima suppression of the Harris responses within a small radius around each extracted keypoint. If corresponding keypoints on the two volumes are above a certain threshold, we consider them a suitable patch pair and use it for pre-training of our network.

The goal of our method is to produce a local feature encoding that maps the local neighborhood of an object around a 3D keypoint on a 3D surface to a vector while being invariant to local changes around the object of interest. We learn this change-invariant descriptor by using the object alignments and sampling dynamic patches from our proposed 3RScan-dataset. So, once converged we fine-tune the static network with dynamic 3D patches specifically generated around points of interest on moving objects. To learn only higher level features, during fine-tuning, we freeze the first layers and only train the multi-scale encoder branch

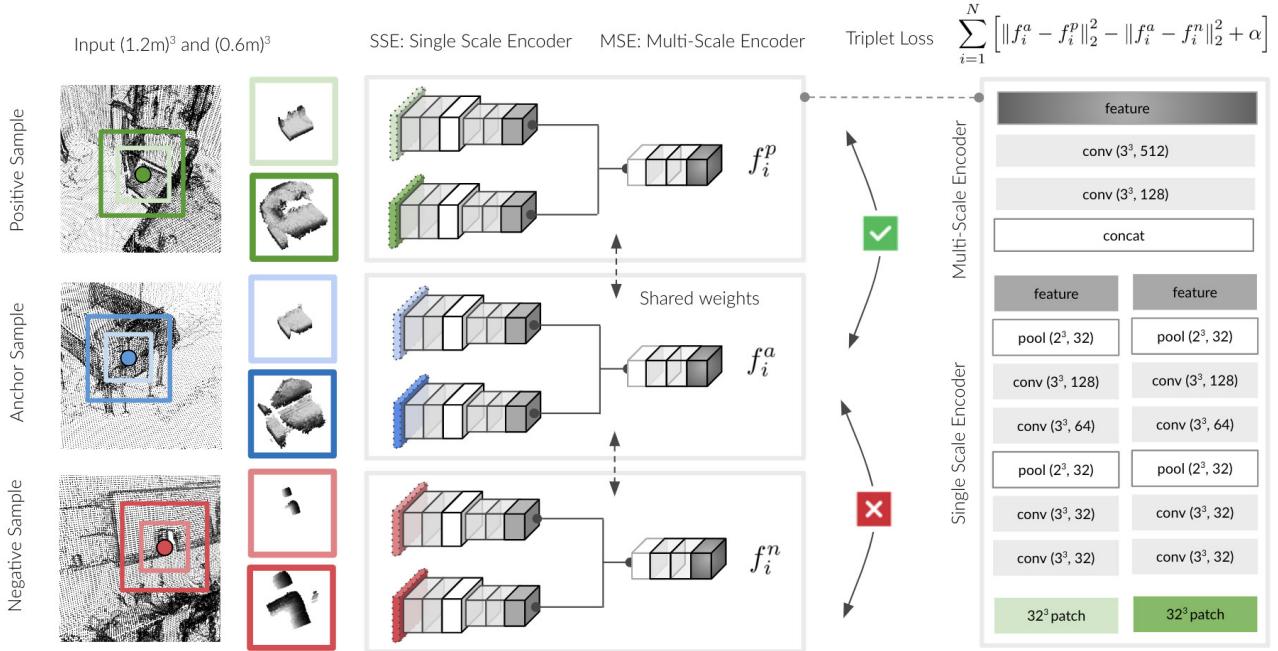


Figure 6: Our multi-scale triplet network architecture: during training, each anchor (blue) is paired with a positive (green) and a negative sample (red). The network minimizes the distance between the positive samples and maximizes negative sample distances by processing two scales in a separate branch each. Weights are shared in each SSE block of the same size and the MSE.

of our network. Correspondence pairs are generated in a self-supervised fashion while using the ground truth pose annotations of our training set to find high keypoint responses in the same small radius around each source 3D keypoint. The negative counterpart of each triplet is randomly selected from another training scene but also includes TSDF patches on removed objects. Random rotation augmentation is applied to enlarge our training data.

4.5. 6DoF Pose Alignment

To re-localize object instances, we first compute features for keypoints on the source objects and the whole target scene. Correspondences for the model keypoints are then found via k-nearest neighbour search in the latent space of the feature encoding of the points in the scene. After outliers are filtered with RANSAC, remaining correspondences serve as an input of a 6DoF pose optimization. Given the remaining two sets of correspondences on the source object $\mathbf{O} = p_1, p_2, \dots, p_n \in \mathbb{R}^3$ and the target scene $\mathbf{S} = q_1, q_2, \dots, q_n \in \mathbb{R}^3$ we then want to find an optimal rigid transformation that aligns the two sets. Specifically, we want to find a rotation \mathbf{R} and a translation \mathbf{t} such that

$$(\mathbf{R}, \mathbf{t}) = \arg \min_{\mathbf{R} \in SO(d), \mathbf{t} \in \mathbb{R}^3} \sum_{i=1}^n \|(\mathbf{R}p_i + \mathbf{t}) - q_i\|^2. \quad (3)$$

We solve this optimization using Singular Value Decomposition (SVD). The resulting 6DoF transformation gives us a pose that aligns the model to the scene. Qualitative results of our alignment method with corresponding ground truth alignments on some scans of our 3RScan-dataset are shown in Figure 7.

5. Evaluation

In the following, we show quantitative experimental results of our method by evaluating it on our newly created 3RScan-dataset. In the first section, we compare the ability of different methods to match dynamic patches around keypoints on annotated changed objects. Our proposed multi-scale network is then evaluated on the newly-created benchmark for re-localization of object instances.

5.1. Correspondence Matching

For accurate 6D pose estimation in changing environments, robust correspondence matching is crucial. The feature matching accuracy of different network architectures is reported in Table 3. Each network is pre-trained with static samples (marked as *static*) and then fine-tuned on dynamic patches from the training (marked as *dynamic*). The F1 score, accuracy, precision, false positive rate (FPR) and error rate (ER) at 95% recall are listed and visualized with



Figure 7: Qualitative results of 3D rigid object instance re-localization (RIO) of our learned multi-scale method in different changing environments. Different instances, taken from the reference scan, are visualized with different colors on top of the re-scan.

Table 3: Evaluation: keypoint matching of dynamic 3D TSDF patches in 3RScan at 95% recall.

Method (train)	F1	Accuracy	Precision	FPR	ER	Top-1	Top-3	Top-5	Top-10
RIO-singlescale 60cm (static)	71.54	62.21	57.37	70.60	75.59	2.17	4.12	5.96	17.56
RIO-singlescale 120cm (static)	74.17	66.92	60.83	61.18	66.16	3.94	4.58	8.21	20.38
RIO-singlescale 120cm (dynamic)	78.71	74.29	67.17	46.43	51.41	6.26	7.26	9.58	27.82
RIO-multiscale (static)	85.58	83.98	77.82	27.09	32.04	30.73	53.48	69.61	89.03
RIO-multiscale (dynamic)	94.37	94.33	93.61	6.50	11.35	64.10	86.20	93.40	98.30

their respective PRC graphs in Figure 8. Additionally to the 1:1 matching accuracy, we also use a top-1 metric: the percentage of top-1 placements of a positive patch given 50 randomly chosen negative patches. Such a metric better represents the real test case of object instance re-localization where several negative samples are compared against a positive keypoint. It can be seen that our multiscale network architecture – even if only trained with static data – outperforms all single scale architectures by a large margin and improves further to an *F1* score of 94.37 if additionally trained with dynamic data.

5.2. Object Instance Re-localization

In the following we discuss results on our newly created benchmark that has been carried out on the test set of 3RScan which is provided with the data. We evaluate our method against hand-crafted features from PCL [23] such as SHOT [27] and FPFH [22, 1]. A transformation of each object instance is computed separately by (1) sampling keypoints, (2) extracting descriptors at each keypoint followed by a (3) correspondence matching and (4) RANSAC-based filtering. A learned baseline we evaluate against is 3DMatch [32].

Table 4: Performance of object instance re-localization. Numbers are reported in terms of average % correct rotation and translation predictions. MTE (Median Translation Error) is measured in meters while the MRE (Median Rotation Error) is in degrees

Method (train)	Recall <0.1m, 10°	MRE [deg]	MTE [m]	Recall <0.2m, 20°	MRE [deg]	MTE [m]
FPFH [22]	2.61	7.25	0.0645	8.36	10.57	0.0776
SHOT [27]	6.79	5.35	0.0268	12.27	8.18	0.0393
3DMatch (dynamic)	5.48	5.81	0.0542	13.05	7.30	0.0708
RIO-multiscale (static)	9.92	4.33	0.0425	17.75	6.39	0.0545
RIO-multiscale (dynamic)	15.14	4.75	0.0437	23.76	6.08	0.0547

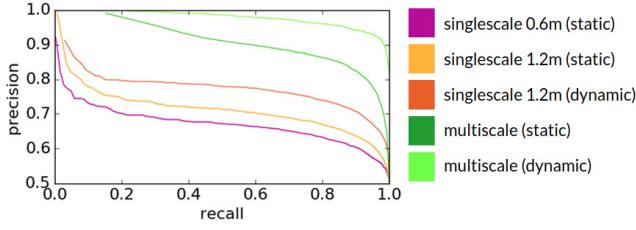


Figure 8: Precision-Recall Curves (PRC) of the dynamic keypoint matching task, corresponding to the different methods evaluated and listed in Table 3.

Table 5: Matching accuracy of the different methods for different instance categories at <0.2m, 20° and our method trained on static (RIO-S) and dynamic data (RIO-D). See supplementary for detailed class description.

class	FPFH	SHOT	3DMatch	RIO-S	RIO-D
seating	5.08	12.71	6.78	14.41	21.19
table	9.33	5.33	21.33	25.33	29.33
items	5.06	13.92	7.59	11.39	16.46
bed / sofa	56.52	21.74	34.78	34.78	47.83
cushion	0.00	15.52	8.62	8.62	10.34
appliances	11.11	16.67	33.33	44.44	55.56
structure	0.00	0.00	8.33	16.67	33.33
avg.	12.44	12.27	17.25	22.23	30.58

It computes a feature given a patch around a keypoint. We trained 3DMatch on $30 \times 30 \times 30$ static positive and negative patches of 30 cm size generated with our dataset as described in the original paper. We evaluate the predicted rotation \mathbf{R}_p and translation \mathbf{t}_p against the ground truth annotation \mathbf{R}_{GT} and \mathbf{t}_{GT} according to equation 5 and 4. An instance has successfully been aligned if the alignment error for the translation \mathbf{t}_Δ and rotation \mathbf{R}_Δ are lower than $t \leq 10$ cm, $r \leq 10^\circ$ or $t \leq 20$ cm, $r \leq 20^\circ$. Please note that respective symmetry are considered in the error computation:

$$\mathbf{t}_\Delta = \mathbf{t}_p - \mathbf{t}_{GT} \quad (4)$$

$$\mathbf{R}_\Delta = \mathbf{R}_p^{-1} \mathbf{R}_{GT} \rightarrow \text{axis angle} \quad (5)$$

Evaluation results for all object instances are listed in Table 4 and Table 5. While classical hand-crafted methods still perform reasonable well – especially for more descriptive objects such as sofas and beds – our method outperforms them with a large margin. Qualitative results are shown in Figure 7.

6. Conclusion

In this work, we release the first large-scale dataset of real-world sequences with temporal discontinuity that consists of multiple scans of the same environment. We believe that the new task of object instance re-localization (RIO) in changing indoor environments is a very challenging and particularly important task, yet to be further explored. Besides 6D object instance alignments in those changing environments, 3RScan comes with a large variety of annotations designed for multiple benchmark tasks including – but not limited to – persistent dense and sparse SLAM, change detection or camera re-localization. We believe that 3RScan helps the development and evaluation of these new algorithms and we are excited to see more work in this domain to, in the end, accomplish persistent, long-term understanding of indoor environments.

Acknowledgment

We would like to thank the volunteers who helped with 3D scanning, all expert annotators, as well as Jürgen Sturm, Tom Funkhouser and Maciej Halber for fruitful discussions. This work was funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), the ERC Starting Grant Scan2CAD (804724), TUM-IAS for the Rudolf Mößbauer Fellowship, and a Google Research and Faculty award.

References

- [1] Aitor Aldoma, Markus Vincze, Nico Blodow, David Gossow, Suat Gedikli, Radu Bogdan Rusu, and Gary Bradski. CAD-model recognition and 6DOF pose estimation using 3D cues. In *International Conference on Computer Vision (ICCV)*, 2011.
- [2] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel Xuan Chang, and Matthias Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [3] Angel Xuan Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017.
- [4] Angela Dai, Angel Xuan Chang, Manolis Savva, Maciej Halber, Tom Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [5] Angela Dai and Matthias Nießner. 3DMV: Joint 3D-Multi-View Prediction for 3D Semantic Scene Segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.
- [6] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. In *ACM Transactions on Graphics (TOG)*, 2017.
- [7] Haowen Deng, Tolga Birdal, and Slobodan Ilic. PPF-FoldNet: Unsupervised Learning of Rotation Invariant 3D Local Descriptors. In *European Conference on Computer Vision (ECCV)*, 2018.
- [8] Marius Fehr, Fadri Furrer, Dryanovski Ivan, Jürgen Sturm, Igor Gilitschenski, Roland Siegwart, and Cesar Cadena. TSDF-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In *International Conference on Robotics and Automation (ICRA)*, 2017.
- [9] Michael Firman. RGBD Datasets: Past, Present and Future. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2016.
- [10] Ji Hou, Angela Dai, and Matthias Nießner. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [11] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset. In *British Machine Vision Conference (BMVC)*, 2018.
- [12] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric Object-Level SLAM. In *International Conference on 3D Vision (3DV)*, 2018.
- [13] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks. In *International Conference on Robotics and Automation (ICRA)*, 2017.
- [14] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew Davison. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? In *International Conference on Computer Vision (ICCV)*, 2017.
- [15] Yoshikatsu Nakajima, Keisuke Tateno, Federico Tombari, and Hideo Saito. Fast and Accurate Semantic Mapping through Geometric-based Incremental Segmentation. In *International Conference on Intelligent Robots and Systems (IROS)*, 2018.
- [16] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision (ECCV)*, 2012.
- [17] Richard Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.
- [18] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. In *ACM Transactions on Graphics (TOG)*, 2013.
- [19] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [20] Charles Ruizongtai Qi, Li Yi, Hao Su, and Leonidas Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
- [21] Dario Rethage, Johanna Wald, Jürgen Sturm, Nassir Navab, and Federico Tombari. Fully-convolutional point networks for large-scale point clouds. In *European Conference on Computer Vision (ECCV)*, 2018.
- [22] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *International Conference on Robotics and Automation (ICRA)*, 2009.
- [23] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *International Conference on Robotics and Automation (ICRA)*, 2011.
- [24] Renato Salas-Moreno, Richard Newcombe, Hauke Strasdat, Paul Kelly, and Andrew Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [25] Shuran Song, Samuel Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [26] Shuran Song, Fisher Yu, Andy Zeng, Angel Xuan Chang, Manolis Savva, and Thomas Funkhouser. Semantic Scene Completion from a Single Depth Image. In *Conference on Computer Vision and Pattern Recognition*, 2017.

- [27] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique Signatures of Histograms for Local Surface Description. In *European Conference on Computer Vision (ECCV)*, 2010.
- [28] Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In *International Conference on 3D Vision (3DV)*, 2016.
- [29] Julien Valentin, Vibhav Vineet, Ming-Ming Cheng, David Kim, Jamie Shotton, Pushmeet Kohli, Matthias Nießner, Antonio Criminisi, Shahram Izadi, and Philip Torr. SemanticPaint: Interactive 3D Labeling and Learning at Your Fingertips. In *ACM Transactions on Graphics (TOG)*, 2015.
- [30] Johanna Wald, Keisuke Tateno, Jürgen Sturm, Nassir Navab, and Federico Tombari. Real-Time Fully Incremental Scene Understanding on Mobile Platforms. In *International Conference On Intelligent Robots and Systems (IROS)*, 2018.
- [31] Thomas Whelan, Stefan Leutenegger, Renato Salas Moreno, Ben Glocker, and Andrew Davison. ElasticFusion: Dense SLAM Without A Pose Graph. In *Robotics: Science and Systems*, 2015.
- [32] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

7. Supplemental Material

In this supplemental document, we provide additional information about the proposed dataset such as statistics, scene examples and a detailed description about the annotation process.

Dataset

Scanning Interface We tailored a mobile app running on Google Tango with pre-annotation functionality as a scanning interface (see Figure 9). Some users gave lightweight instructions on the scene changes; these instructions served as guidelines later in the annotation process.

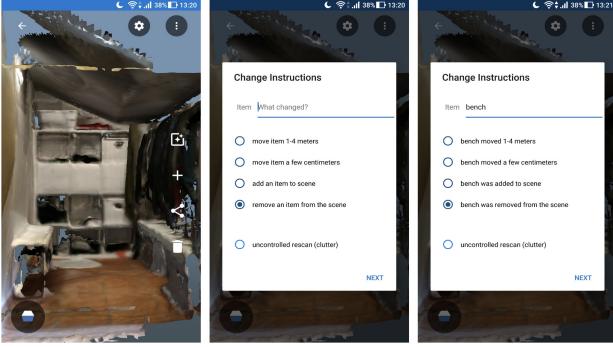


Figure 9: Annotation Interface used for data acquisition.

Scene Matching and Alignment For each uploaded scan, scene candidates are computed. Since a 3D scene matching is expensive scan pairs are found in 2D instead by conducting a similarity search in the texture uv-map of the mesh. These matches are then to be manually adjusted. Once the reference for each scene is assigned, the IMU normalized scans are globally registered via a coarse to fine correspondence-based 2D ICP together with RANSAC and refined with a global 3D ICP. An additional verification as well as an optional manual, keypoint based alignment ensures high quality.

Preprocessing Additionally to a server-side offline processing of the RGB-D sequences that results in texture mapped 3D reconstructions, an offline 3D segmentation is triggered. This 3D segmentation is utilized by the semantic segmentation interface proposed by Dai *et al.* [4]. Further, this 3D segmentation – together with the aforementioned 3D alignment – serves as the basis for the propagation of semantic labels from the references to the re-scans and after a manual clean-up procedure results in the final instance segmentation shown in Figure 17. In the current snapshot of the dataset almost all of our scans have an instance segmentation coverage of above 90% (see Figure 10) with an

average scene coverage > 98%. In total, 48k instances are annotated with 534 unique labels.

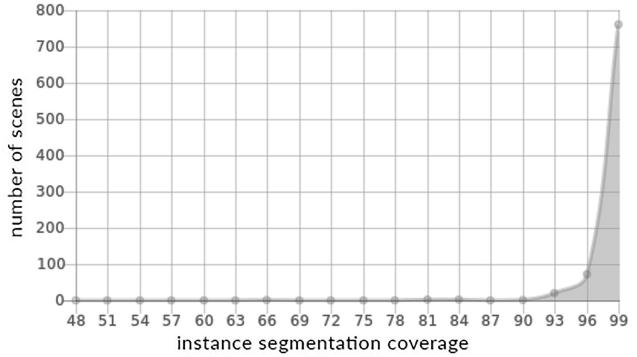


Figure 10: Number of scans with corresponding instance annotation coverage.

RGB-D sequences Our dataset consists of around 363k calibrated RGB-D and depth images. Since raw RGB and depth sequences from Tango are of varying frame rates and spatial resolution, a spatial and temporal calibration procedure of the raw images is applied. Further, to remove rectification lines present in Google Tango depth images, a median filter is used before the spatial calibration of the images. Camera trajectories for SLAM are visualized in Figure 18 and since global scene-to-scene mappings are provided these can easily be transferred into the same coordinate systems (see last two rows of Figure 18). Further, we also show 2D projections of our textured 3D models with aligned camera poses in Figure 14.

Scene Type Further, instead of assigning a $1 - n$ -relationship of different room types per scene, our 3D reconstructions are annotated with m corresponding scene functionalities (sleeping, eating, working, etc.) in an $n - m$ fashion. This shows the high variety of scenes in 3RScan.

Instance Change Annotation The annotation interface for annotating instance changes is a web-based tool (Figure 16) where each scene is rendered next to its corresponding reference. When an object is selected in the re-scan (see green dot) its instance segmentation from the reference scan is automatically segmented. Please note, that this requires the instance IDs to be consistent across scans of the same environments. Hovering over the objects gives shows the label and the ID of the instance and allows to potentially fix the underlying semantic segmentation. In the alignment view this instance is then shown next to the re-scan such that corresponding keypoints can easily be selected. Once

enough keypoints are set, a Procrustes based alignment (Kabsch algorithm) is triggered that computes a transformation that aligns the object to the scene. For non-rigid changes and removed as well as added objects the instances IDs are tracked.

Symmetry A subset of the changed objects in the dataset are symmetric. We follow the symmetry annotation described in Avetisyan *et al.* [2] and categorize each object's rotational symmetry around the canonical axis to the classes C_2 , C_4 and C_∞ . 22% of the objects have a symmetry as listed in Table 6. We take this into account when evaluating the predictions against ground truth poses.

Table 6: Symmetry properties of the instances in 3RScan

symmetry	none	C_2	C_4	C_∞
# scans	1513	220	82	132

Statistics A focus during data acquisition was the capturing of a variety of realistic scene changes in controlled and uncontrolled environments over a time span of more than 12 months. The number of references scenes with re-scanning frequencies are plotted in Figure 11. Further, 3289 instance transformations – of 1947 different objects – are provided with the data. But since the transformations give the object pose from the reference to one re-scan the alignment for another re-scan can easily be computed. For evaluation these changed object categories are mapped to 9 different classes as listed in 7.

Table 7: Description of instance mapping used in the evaluation

class	description
seating	different chairs, stools, benches
table / cabinet	different tables, commode, shelves
bed / sofa	upholstery, sofas, beds
appliances	appliances, sanitary equipment
cushions	pillows, bean bags, ottoman
items	small and portable items, boxes
structure	windows, doors

These changed object instances are labelled with 187 different categories. The majority of instances include movements of objects and more portable furniture items such as chairs, pillows, boxes or smaller tables. Naturally, these objects involve most human interaction. Figure 13 gives an overview of the motion of these annotated objects. However, we also annotated objects that slightly change their appear-

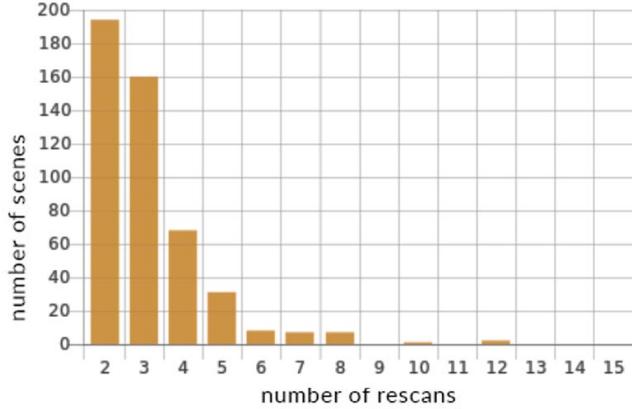


Figure 11: Number of scenes vs. number of re-scans.

ance over time such as toilets. Detailed statistics are given in Figure 12.

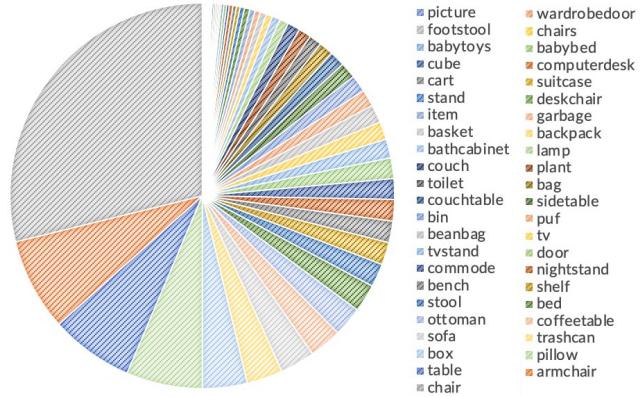


Figure 12: Object statistics.

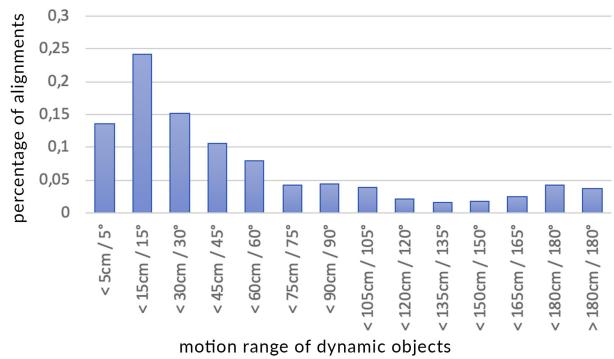


Figure 13: Histogram of object instance alignment binned using respective transformation or rotation change.

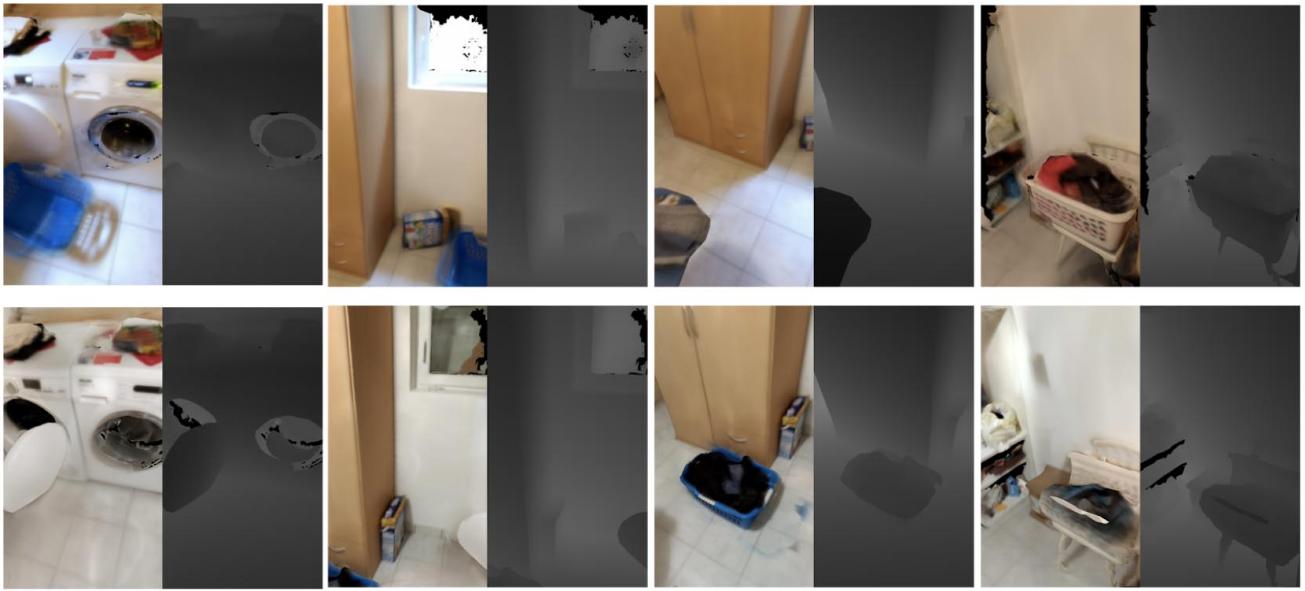


Figure 14: Example 2D projections of the color and depth of two corresponding reconstructions with natural scene changes.



Figure 15: List of mutually inclusive scene functionalities with corresponding visual examples, from top left to bottom right: (a) working, (b) sleeping, (c) eating, (d) entertainment, (e) seating, (f) storage, (g) reading, (h) food preparation, (i) cleaning and (j) personal hygiene.

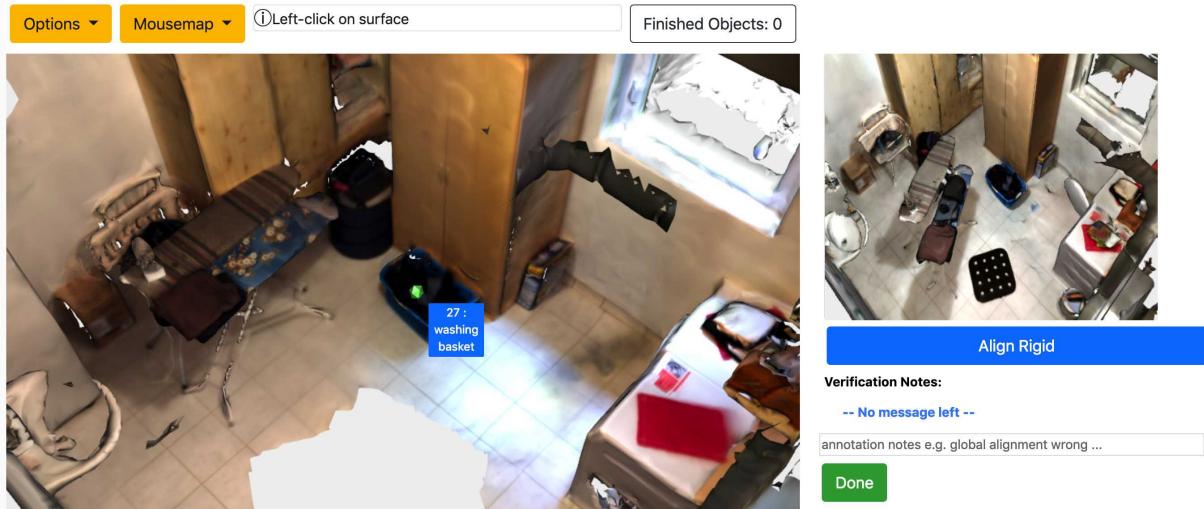


Figure 16: Instance Change Annotation Tool: Overview and selection view of the instance alignment annotation.



Figure 17: Visualization of different annotated instances in scans of 3RScan.



Figure 18: SLAM: Different 3D Scenes with Camera Trajectories in green used for training and generation of the static TSDF samples.