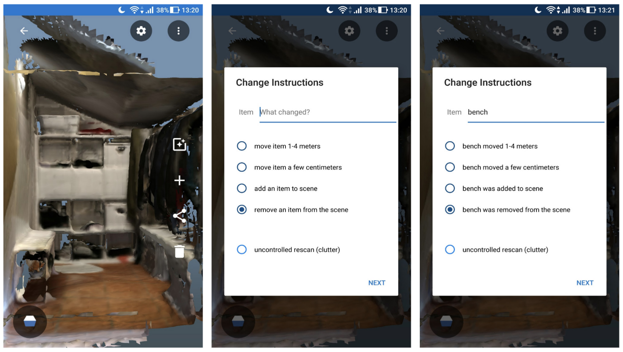
# Supplemental Material
# RIO: 3D Object Instance Re-Localization in Changing Indoor Environments

Johanna Wald [1] Armen Avetisyan [1] Nassir Navab [1] Federico Tombari [1,2,*] Matthias Nießner [1,*]
[1] Technical University of Munich [2] Google

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 1). Some users gave lightweight instructions on the scene changes; these instructions served as guidelines later in the annotation process.



Change Instructions

Item What changed?

○ move item 1-4 meters
○ move item a few centimeters
○ add an item to scene
● remove an item from the scene

○ uncontrolled rescan (clutter)

Change Instructions

Item bench

○ bench moved 1-4 meters
○ bench moved a few centimeters
○ bench was added to scene
● bench was removed from the scene

○ uncontrolled rescan (clutter)

**Figure 1:** 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 conducing 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*. [2]. 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 9. In the current snapshot of the dataset almost all of our scans have an instance segmentation coverage of above 90% (see Figure 3) with an average scene coverage $> 98\%$. In total, 48k instances are annotated with 534 unique labels.

**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 10 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 10). Further, we also show 2D projections of our textured 3D models with aligned camera poses in Figure 7.

**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 8) 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

[*] Authors share senior authorship.

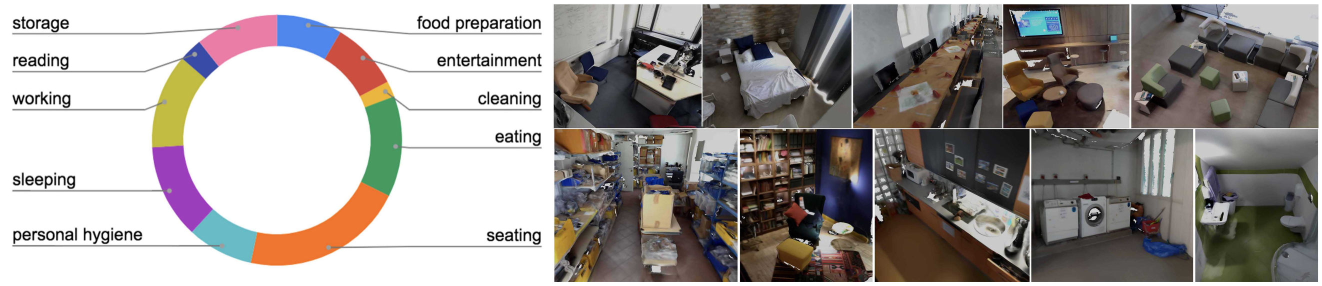


**Figure 2:** List of mutually inclusive scene functionalities with corresponding visual examples, from top left to bottom right: (a) working, (b) sleeping, (c) eating, (d) entertainment, (e) seating, (f) storage, (g) reading, (h) food preparation, (i) cleaning and (j) personal hygiene.

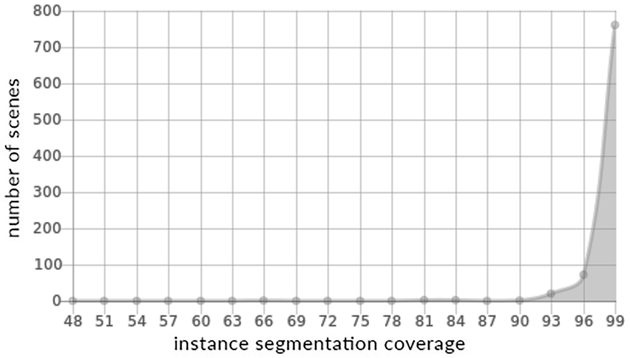


**Figure 3:** Number of scans with corresponding instance annotation coverage.

**Table 1:** Symmetry properties of the instances in 3RScan

| symmetry | none | $C_2$ | $C_4$ | $C_\infty$ |
|---|---|---|---|---|
| # scans | 1513 | 220 | 82 | 132 |

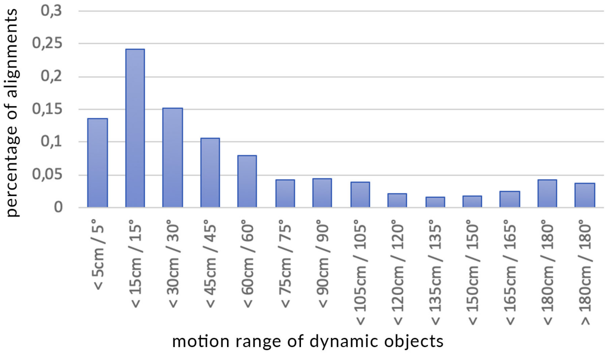


**Figure 4:** Histogram of object instance alignment binned using respective transformation or rotation change.

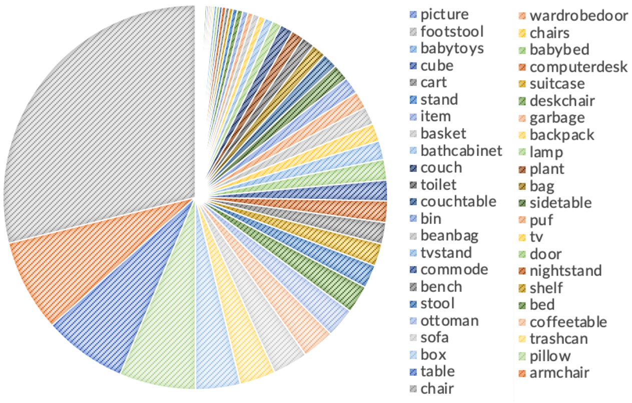


**Figure 5:** Object statistics.

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*. [1] and categorize each object's rotational symmetry around the canonical axis to the classes $C_2$, $C_4$ and $C_\infty$. 22% of the objects have a symmetry as listed in Table 1. We take this into account when evaluating the predictions against ground truth poses.

**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 6. 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 7 different classes as listed in 2.

**Table 2:** 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 |

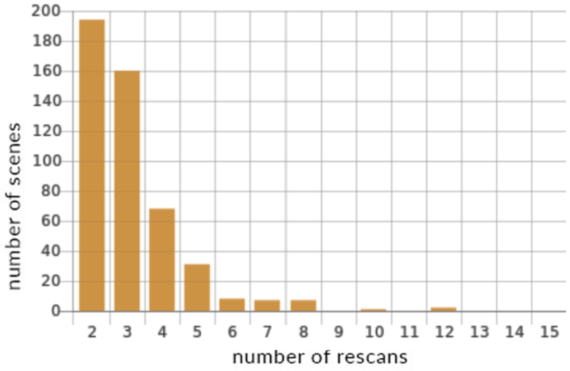


**Figure 6:** Number of scenes vs. number of re-scans.

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 4 gives an overview of the motion of these annotated objects. However, we also annotated objects that slightly change their appearance over time such as toilets. Detailed statistics are given in Figure 5.

## References

[1] 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.

[2] 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.

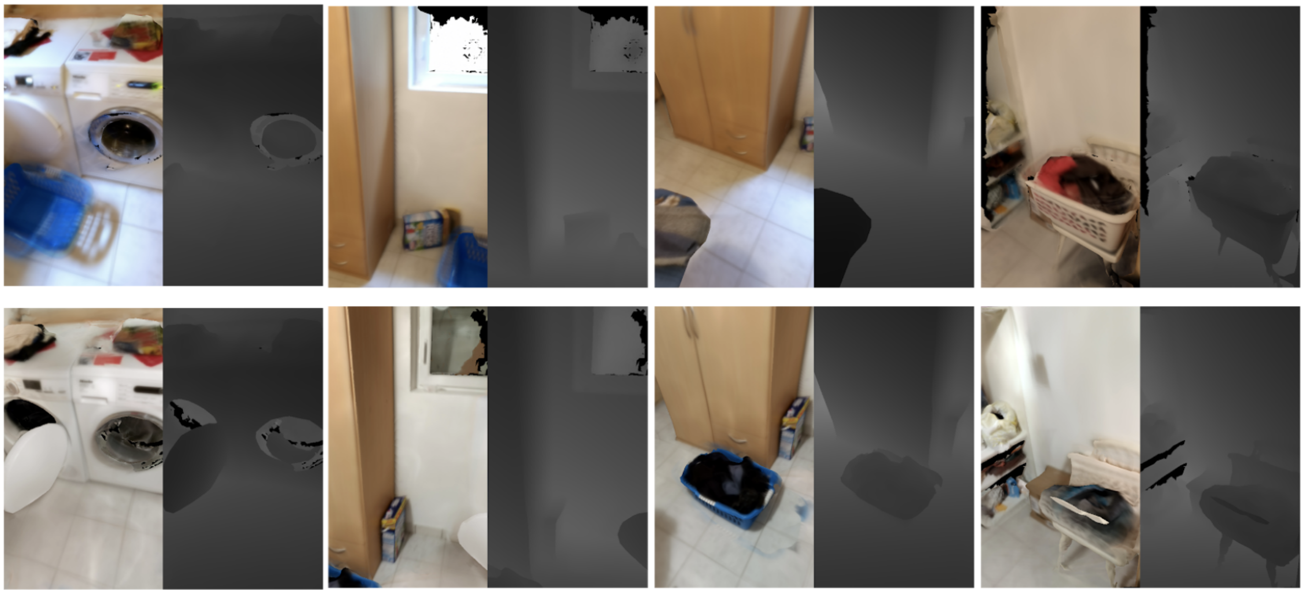
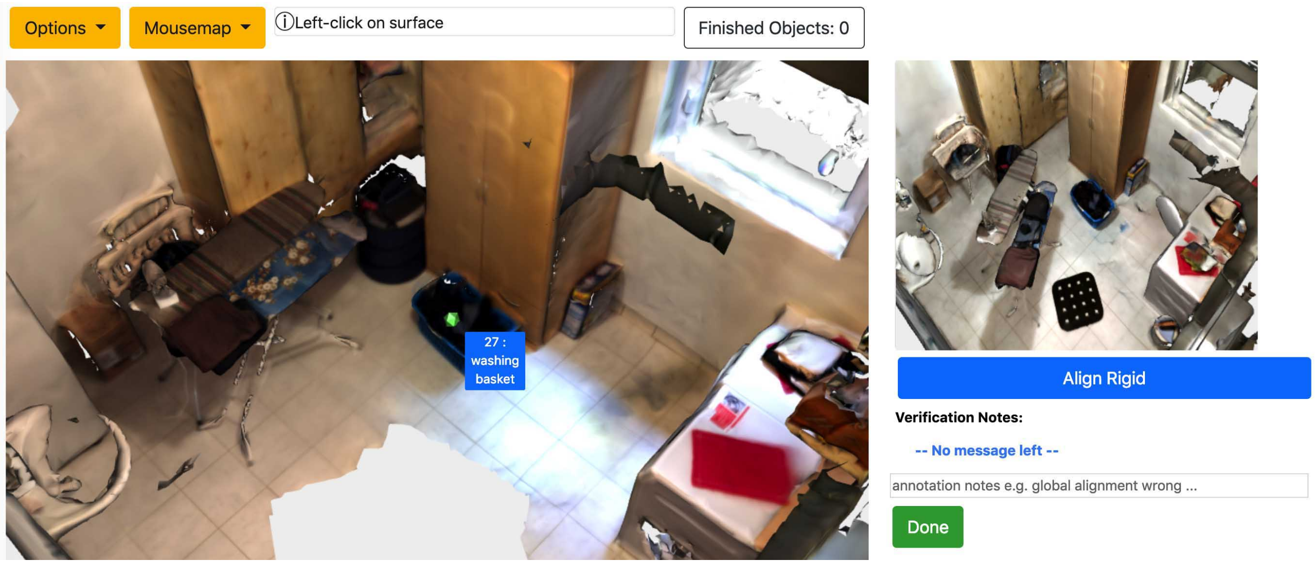


**Figure 7:** Example 2D projections of the color and depth of two corresponding reconstructions with natural scene changes.



Verification Notes:

-- No message left --

annotation notes e.g. global alignment wrong ...

**Figure 8:** Instance Change Annotation Tool: Overview and selection view of the instance alignment annotation.

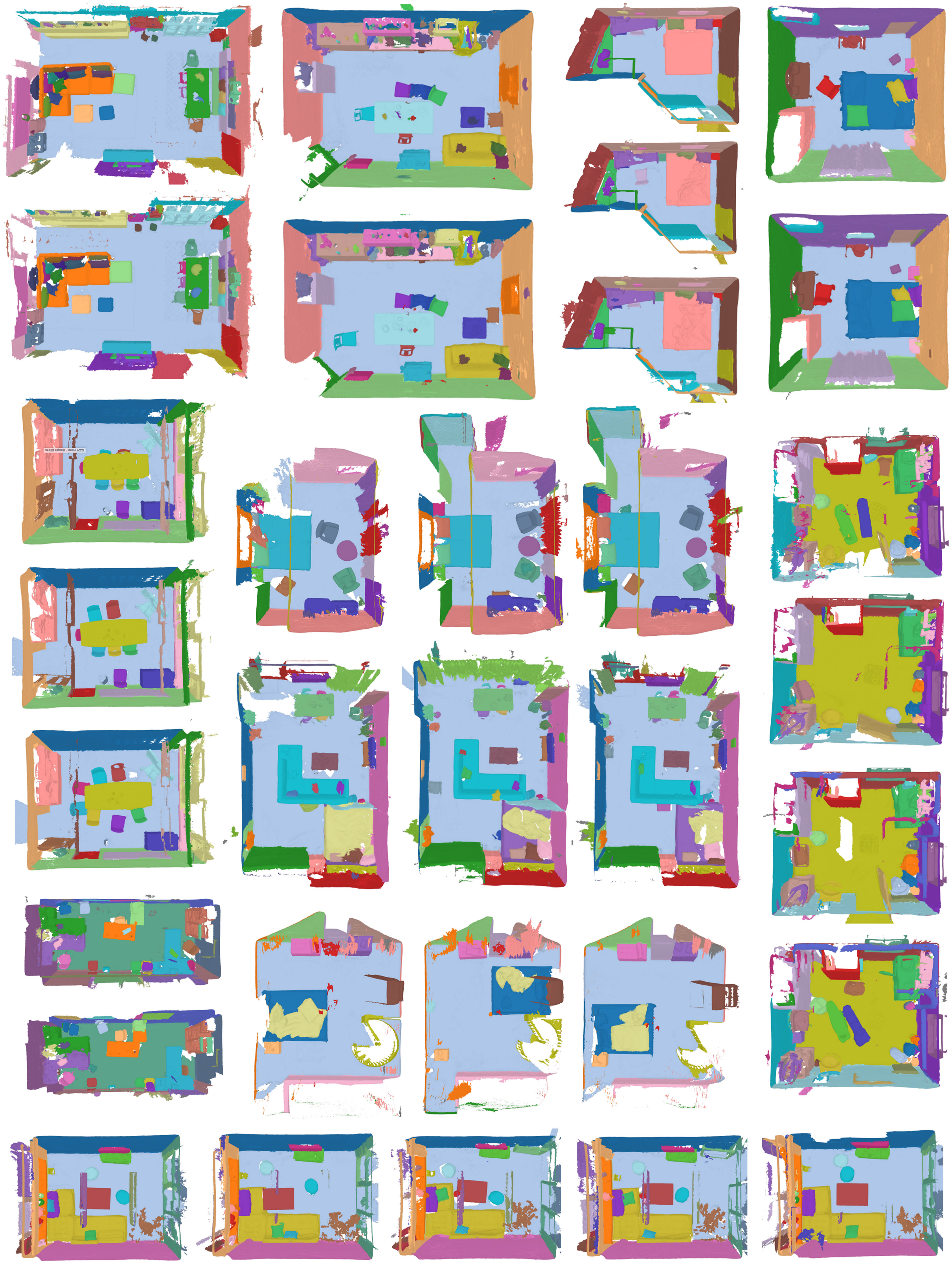


**Figure 9:** Visualization of different annotated instances in scans of 3RScan.

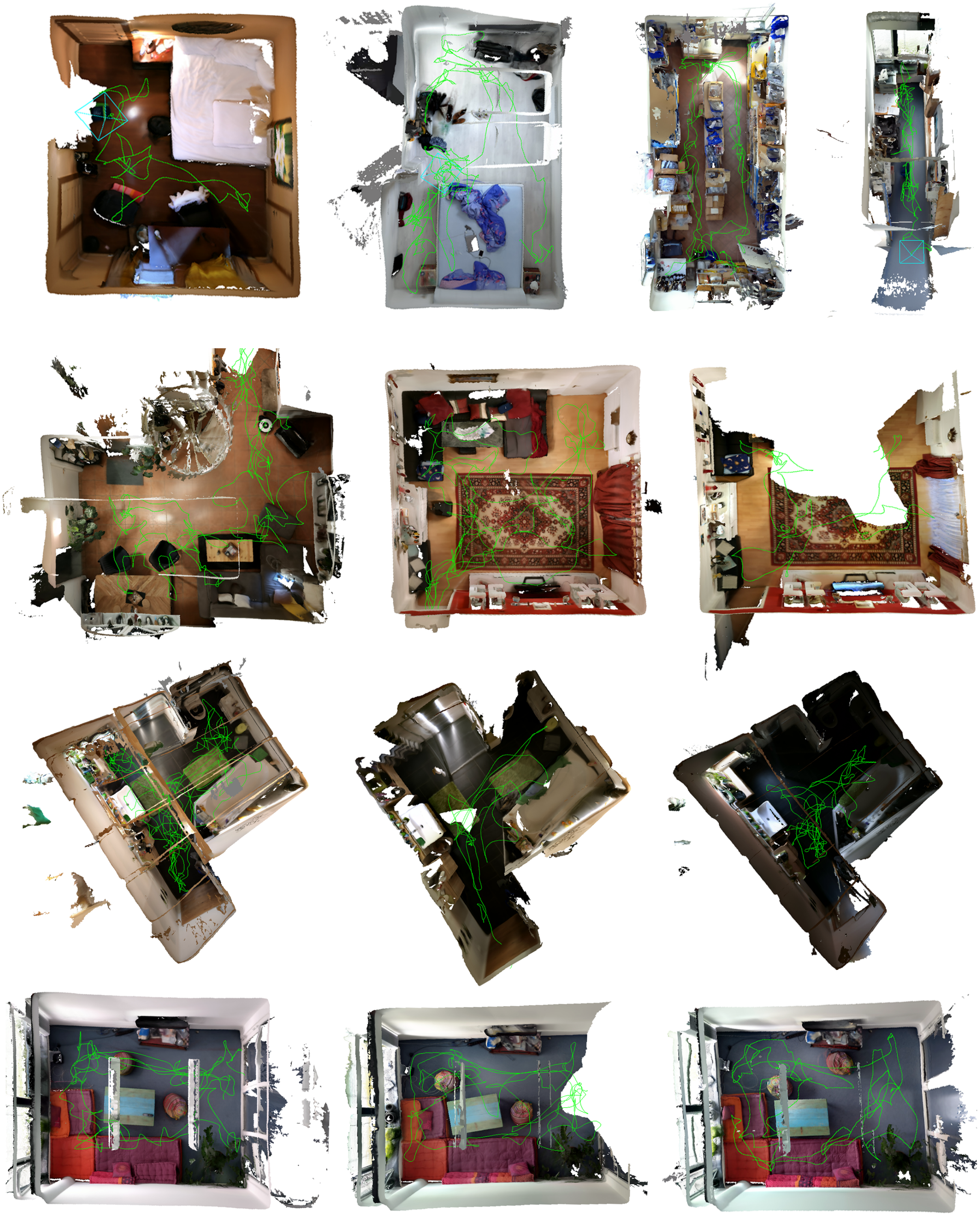


**Figure 10:** SLAM: Different 3D Scenes with Camera Trajectories in green used for training and generation of the static TSDF samples.