# Rescan: Inductive Instance Segmentation for Indoor RGBD Scans
# Supplemental Material

Maciej Halber Yifei Shi Kai Xu Thomas Funkhouser
Princeton University

This document contains supplemental material describing details of the algorithms, datasets, and evaluations for RESCAN.

## 1. Algorithmic Details

In this section, we provide further details pertaining to the energy function and search algorithm.

### 1.1. Hysteresis Term

In this section we provide additional details regarding the *Hysteresis Term*. As seen in the figure 1, coffee tables, trash bins and miscellanea are the most frequently moving objects. A simple common-sense explanation of this behaviour is that smaller objects are likely to be light and easy to move. To incorporate this observation into our objective function, we add a heuristic for selecting value of $\sigma$ in the *Hysteresis Term*. To reiterate, the *Hysteresis Term* is computed as:

$$h + (1-h)exp(\frac{-||T(c_k,i) - T(c_k,j)||_2}{2\sigma^2})$$

We make $\sigma$ dependent on the object $o_k$. The value of $\sigma$ is computed as $\sigma(o_k) = aexp(-bV(o_k)) + c$. $V(o_k)$ returns the approximate volume of object $o_k$ based on its oriented bounding box. Parameters $a, b, c$ were fit so that the resulting $\sigma(o_k)$ is inversely proportional to $V(o_k)$ - larger objects obtain smaller $\sigma$, leading to a smaller *Hysteresis Term* value when object moves significantly. Smaller objects obtain larger $\sigma$, allowing them to move more freely within the scene with less penalty.

### 1.2. Parameters

Table 1 lists all parameters used in our system. Please note that the same set of parameters is used for all of our experiments.

| Parameter | Value |
|---|---|
| Pose Proposal - x-step | 0.1m |
| Pose Proposal - y-step | 0.1m |
| Pose Proposal - $\theta$-step | $\frac{\pi}{10}$ |
| Pose Proposal - Acceptance thresholds $\mathbf{t}$ | {0.5, 0.4, 0.35, 0.25} |
| Objective - Weights $\mathbf{w}$ | {2.0, 0.3, 1.0, 1.8} |
| Objective - Intersection Term $\sigma$ | 0.3 |

Table 1: Parameter settings.

## 2. Rescan Dataset Details

For the purpose of evaluating algorithms on our new task, we have collected a dataset of temporally varying scenes. The Rescan Dataset contains 3D reconstructions of common spaces like lounges, study areas and living rooms. The captured spaces are relatively large, with an average approximate area of $67.58m^2$. The RGBD sequences were captured using the Structure Occipital Sensor, using the Scannet Capture App [2]. 3D reconstructions were obtained using an algorithm described in [3].

| Class | Total Count | Unique Count |
|---|---|---|
| chair | 576 | 184 |
| other | 199 | 72 |
| table | 145 | 40 |
| desk | 46 | 16 |
| sofa | 38 | 12 |
| shelves | 14 | 5 |
| bookshelf | 3 | 1 |
| Sum | 1021 | 330 |

Table 2: Total count describes number of objects in all of scenes in our dataset. Unique count specifies the number of unique instances that have appeared across time.

Each scene was captured between three to five times. Between each capture, the objects within the scene were moved in the way that long-term changes that are likely to occur in such spaces. In most cases, objects were newly introduced or removed from the scene as they would be in natural use.

In total Rescan dataset contains 45 sequences, distributed among 13 distinct scenes (see table 4). For each of the 45 3D reconstructions, we provide semantic instance segmentations describing the contents of the scene (see table

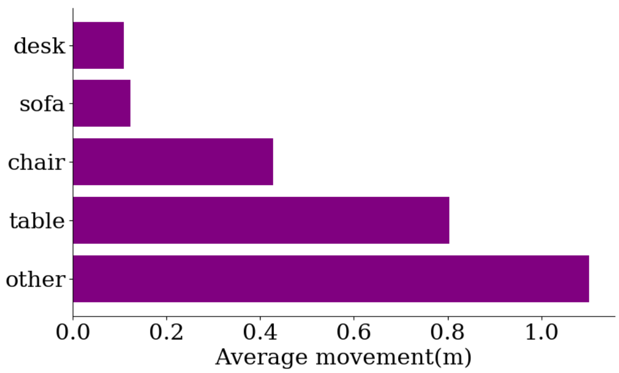



Figure 1: Average movement for objects in the Rescan dataset. We can observe that objects like desk and sofa move infrequently. Chairs move significantly more, but on average stayed relatively close to their previous position. $Table$ and $Other$ categories encompass objects like coffee tables and small trash bin. Such objects were easy to manipulate due to their weight and size. Hence we observe much more motion for objects falling into either of these categories.

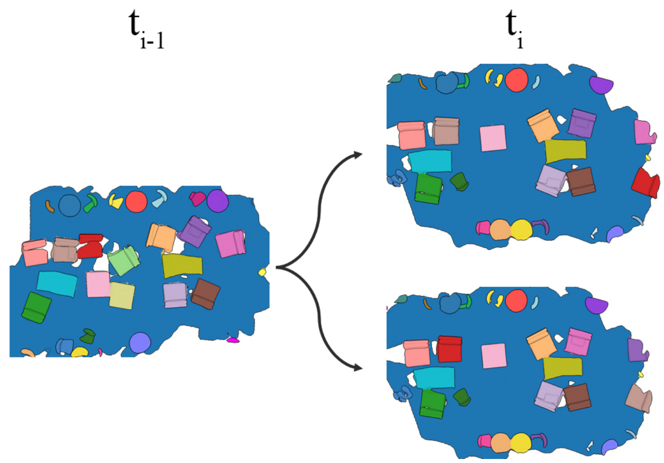



Figure 2: Given an arrangement of objects at $t_{i-1}$, it is often the case that multiple arrangements at $t_i$ make sense. Rescan dataset provides permutation of object id assignment to account for such cases.

2). The main difference between Rescan dataset and other RGBD datasets [4, 2, 1] is the presence of object associations across time. These associations are expressed in the form of stable instance segmentation – object A has the same unique instance id in all time-steps $t_0, t_1, ..., t_n$. Object associations allow us to evaluate algorithms on our task, and they provide additional information about the scenes, like the average distance objects within a certain category have moved (see figure 1).

Since the dataset samples time at very sparse intervals, it includes cases where the correct associations of object instances across time are ambiguous - multiple arrangements of objects are equally likely (see figure 2). To deal with this issue, the Rescan dataset also provides hand-crafted permutations of object instance ids assignments – i.e., sets of object associations that are ambiguous – and it includes evaluation metrics that account for these ambiguities. This allows us to evaluate algorithms without penalties for producing one of many equally likely solutions.

# 3. Evaluation Details

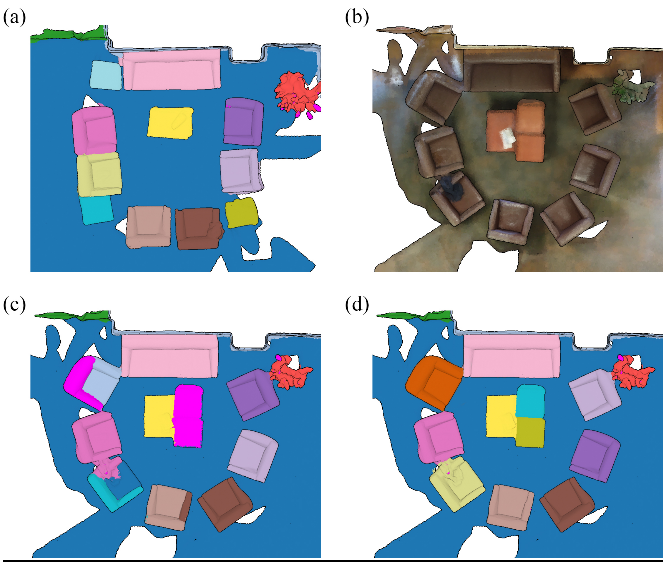
In addition to the comparisons and ablation studies provided in the main paper, we report here results of further ablation studies aimed at characterizing the limits and behavior of our algorithm.

## 3.1. Limited Movement Study

Our first study analyzes the importance of allowing the pose proposal step to search an exhaustive range of possible object placements vs. a heuristic that considers only limited movements. In the limited movement variant of our method we do not perform the dense search for arbitrary poses for each object. Instead we only allow for movement within a $20cm$ radius around the position of the object in the previous arrangements. From results in fig. 3 we can see that the limited movement leads to a significant decrease in performance. For the instance segmentation tasks, results are on average similar, however both method (full movement and limited movement) have different modes of failure. Full movement might produce incorrect permutation of chairs around the table, when *Coverage Term* outweighs the *Hysteresis Term* (as discussed in the main paper). Limited movement does not have that issue, as no additional poses for such chairs were produced. It however misses objects that have moved significantly.

## 3.2. Temporal Stability Study

In our second experiment, we investigate the stability of our method when the number of scene captures grows. To this end we introduce an additional scene containing separate captures at 15 different times. Although this new scene is relatively small, it contains significant object motion and frequent object entrance and exit events, and it spans a long time interval. So, it allows us to more closely investigate characterize our algorithm's behavior in this setting. Figure 4 visualizes the transfer results over all 15 time-steps. In the beginning the method is given ground-truth segmentation, that is used to estimate segmentation for the next time-step and so on. We can see that even with longer sequences our method is able to provide high quality results, and able to estimate correct number of objects even




| Method | Semantic Label | Semantic Instance | Temporal Instance |
|---|---|---|---|
| Limited Movement | 0.747 | 0.675 | 0.623 |
| Full Movement | 0.859 | 0.837 | 0.650 |

Figure 3: A comparison of full vs limited movement schemes. The limited movement variant leads to significant drop in performance. In qualitative comparison (a) is the source scene with instance segmentation. (b) is the target scene visualization. (c) and (d) show result using our method using limited movement and result using full version of the presented method, respectively.

as some exit or enter the scene.

### 3.3. Timings

In this section, we present average running times for all the stages of the proposed pipeline.

| Procedure | Timing(s.) |
|---|---|
| Pose Proposal | 1397.9 |
| Greedy Initialization | 70.2 |
| Simulated Annealing | 93.6 |
| Segmentation Transfer | 45.9 |

Even with the simplifications proposed the Pose Proposal stage is still the most time consuming stage of our system. Our algorithm can be however made easily parallel to produce sets of pose proposals for multiple objects at the same time. Another direction for future work would be to learn conditional embedding of a scene and candidate object to a space that returns correct poses, instead doing purely geometrical matching. However, since this is an offline system, the current run times do not contribute significantly to its overall utility.

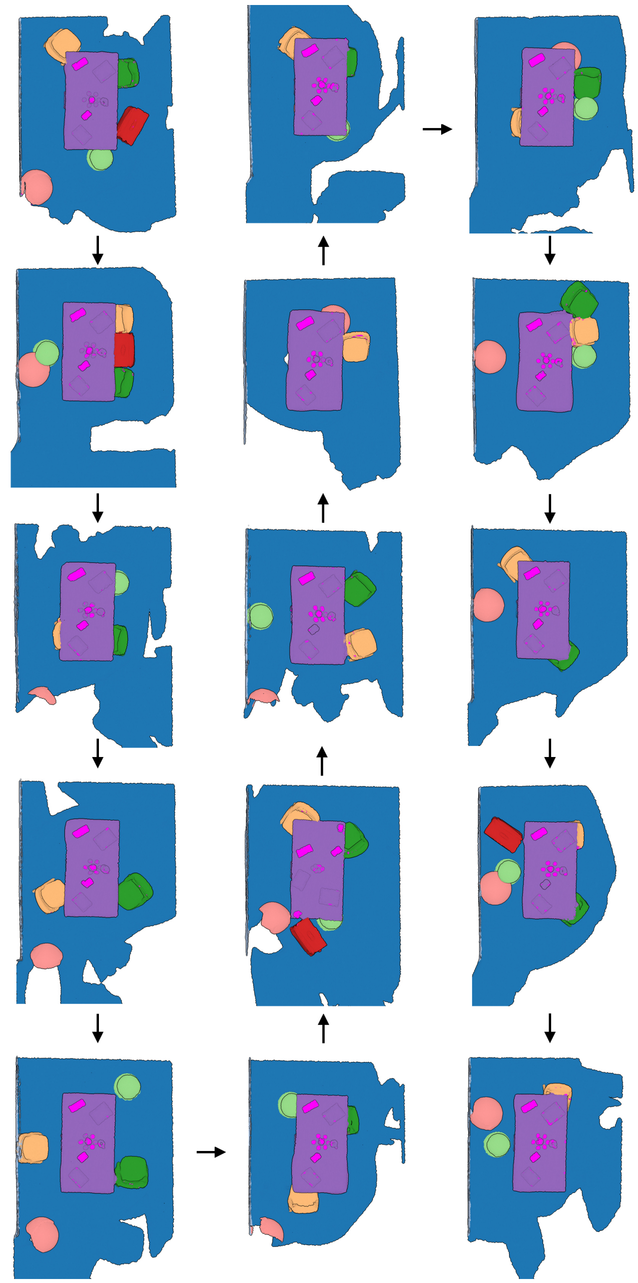


Figure 4: Segmentation transfer result over 15 time-steps. Our algorithm is able to provide stable segmentation transfer over longer sequences.

### 3.4. Failure Cases

In this section, we provide an expanded discussion of the failure modes of our method.

**Novel object's appearance** The quality of the segmentation of the scene at time $t_i$ depends on the amount of information shared between the reconstructions at time $t_{i-1}$ and time

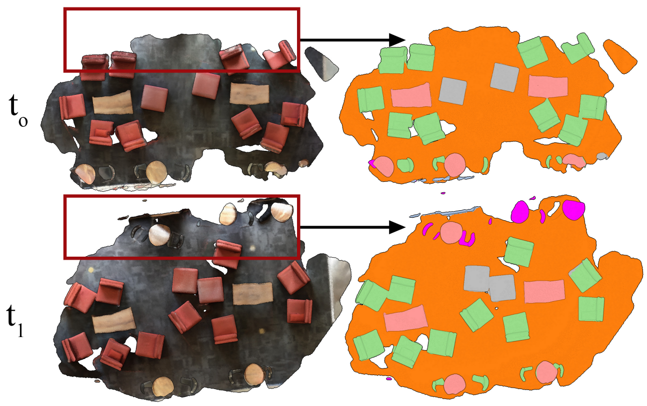
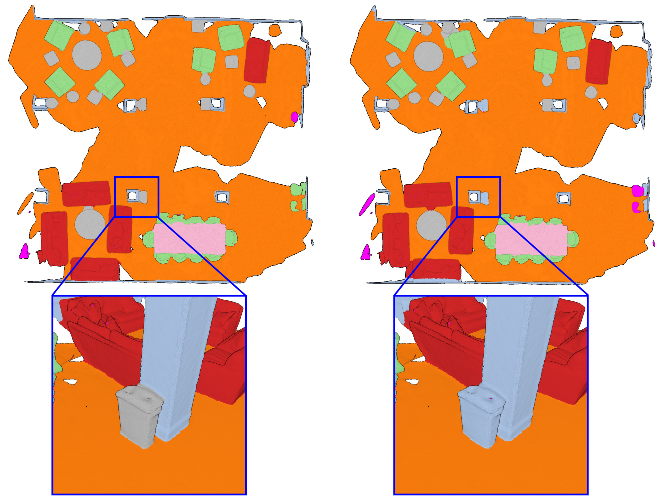



Figure 5: Scene capture at $t_0$ is missing the part scanned in $t_i$(highlighted). Due to the lack of overlap between two captures, some objects at $t_1$ are missed.

$t_i$. If the scene $S_i$ contains significantly more objects than $S_{i-1}$ (either due to the sudden appearance, or an incomplete scanning at $t_{i-1}$), our method prefers to only predict labels for the parts of the scene that are shared between $S_i$ and $S_{i-1}$ (see fig. 5). This is a result of our formulation of the *Hysteresis Term* which aims to create object identity associations between $S_i$ and $S_{i-1}$, and is penalized when adding many novel objects. Secondly, while our method is capable of providing labels for novel objects, this capability is however limited to objects that exist in the temporal model $M_{i-1}$. When a novel object outside $\mathcal{O}$ appears, our method either a) mislabels it or b) leaves it unlabelled. The first case happens if the novel object's geometry is similar to some object $\in M_{i-1}$, so that *Pose Proposal* stage is able to generate potential locations. The second happens if the novel object is completely different than any of the objects in $\mathcal{O}$. In the Rescan dataset we encounter a single case where such situation happen. In figure 6 we can see mislabeled thrash bins - the reason for this is the lack of such objects at previous time-step.

**Small object's movement** Given the form of the *Coverage Term*, larger objects are preferred, as object's contribution to the overall score is proportional to its size. Smaller objects contribute much less, while at the same time are more likely to move. In the section 1.1 we outlined our strategy to combat this issue, however it still occurs in certain cases, like the one visualized in figure 7.

**Partial Geometry** The *Pose Proposal* stage attempts to generate sets of poses for each object $o_k \in \mathcal{O}$ based on the the geometry of the target scene $S_i$. If some parts of the scene were scanned partially, the *Pose Proposal* stage is not be able to find correct candidate poses for such parts of $S_i$. As a result with an incomplete set of poses, the subsequent stage of *Arrangement Optimization* has no hope for success, as correct locations of objects are simply not within its



(a) (b)

Figure 6: (a) Ground truth segmentation. (b) Semantic label prediction. The lack of thrash bins in previous timesteps causes our method to mislabel the thrashbin as a part of the column.

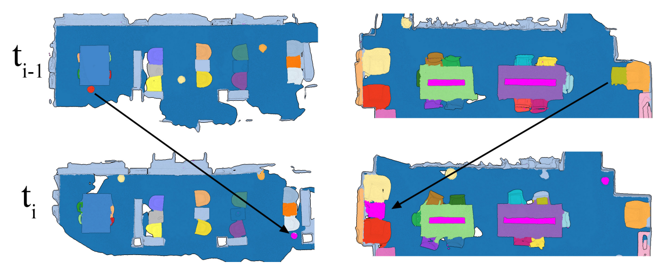



Figure 7: When small objects, like thrash bin or coffee table move significantly, our method may not estimate correct instance labels.

search space. Such cases usually arise at the peripherals of the scene, which were not captured carefully (see fig. 8).

**Permutation of objects** Additional failure case we observe is the permutations of the objects of a similar visual appearance that reside in close spatial proximity. A concrete example of this general concept is "chairs around the table" case. Aforementioned similarity of the objects, combined with a slight movement and imperfections of the reconstruction process causes the Rescan algorithm to confuse the locations of the objects of the same semantic class. This is additionally exacerbated by the fact that that in such settings objects are relatively close, making the *Hysteresis Term* a poor discriminator (figure 9)

Overall the performance of our method is still high despite these issues. Moreover, most of the above issues could be resolved with an alternative formulations of each

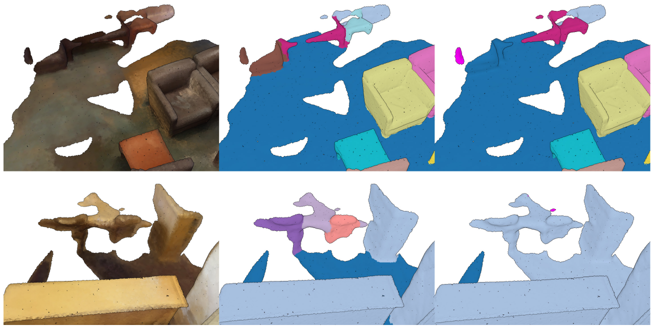


Figure 8: Left - partially scanned geometry. Middle - ground truth segmentation. Right - predicted segmentation. Partial scanning prevents effective pose proposal, resulting in sub-optimal segmentation results.

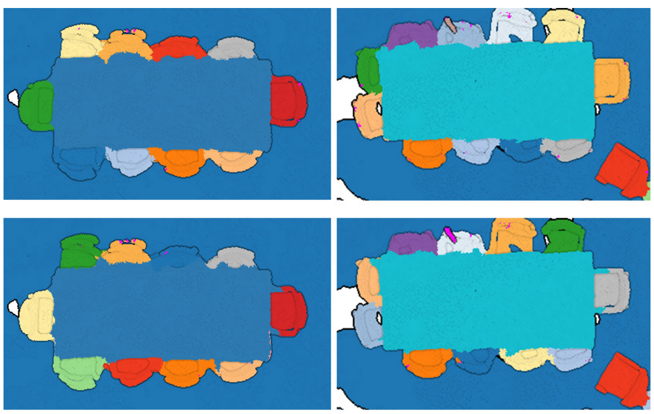


Figure 9: Top - ground truth segmentation. Bottom - predicted segmentation. For sets of visually similar objects our method might produce incorrect permutations of identity assignments.

of the objective function's terms. Learned alternatives for these terms are an interesting direction for the future work. A more fundamental issue comes from the case of an object outside $\mathcal{O}$ appearance. Without a similar object example in the database, our method either mislabels the novel object, or provides no labels. A detection of such cases is also an interesting future direction. One way of resolving it would be to put user in the loop, or combine Rescan algorithm with supervised approaches for instance segmentation.

| $t_0$ | $t_1$ | $t_2$ | $t_3$ | $t_4$ |
|---|---|---|---|---|

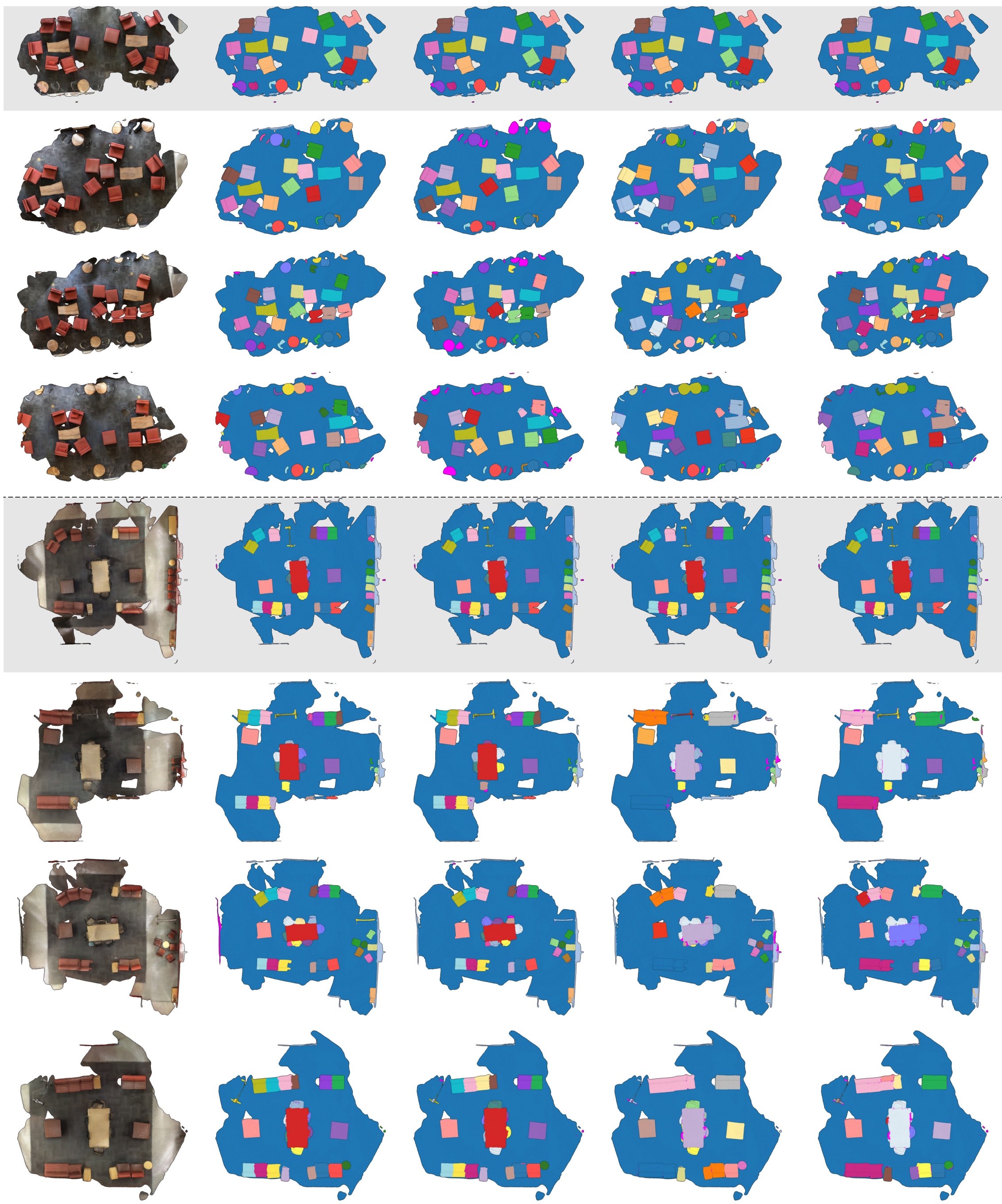
Table 4: Visualization of all scenes in the Rescan Dataset

# A. Qualitative Comparison Results

## A.1. Semantic Instance Transfer Task

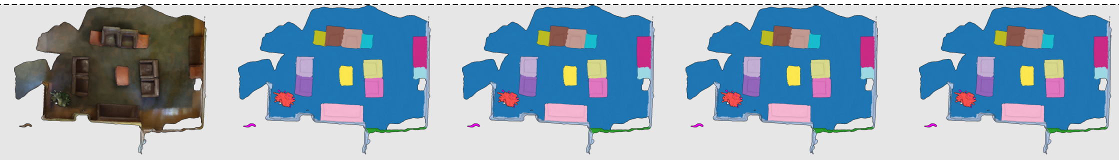
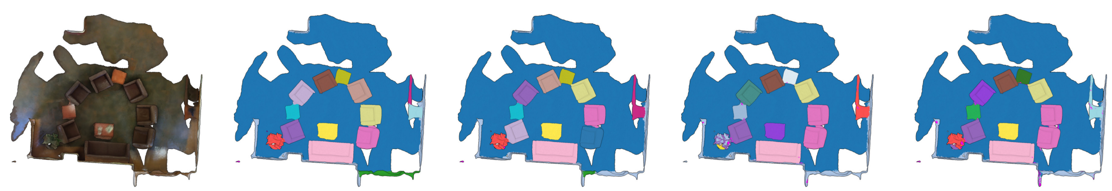
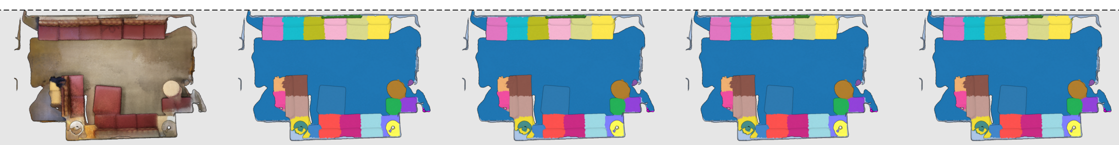
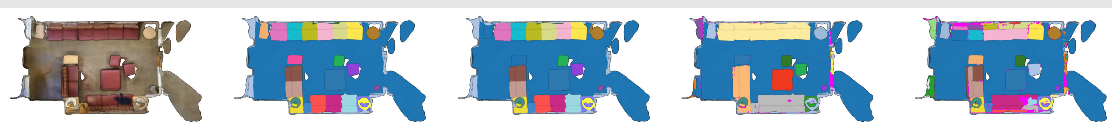
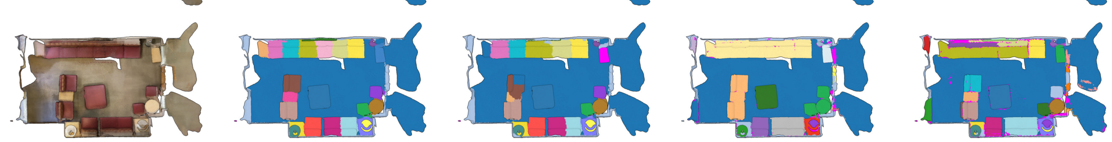
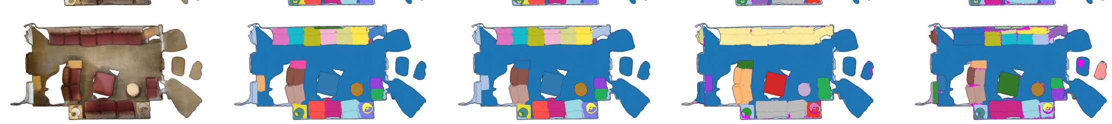
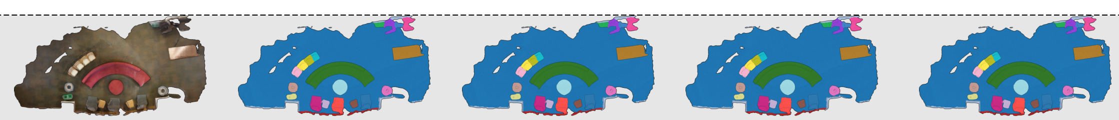
Table 5: *Semantic Instance* and *Instance Transfer Task*. In the first row for each scene (gray background) we visualize ground truth segmentation at time $t_0$. Following rows showcase visualization of instance segmentation estimations. Colors indicate temporal association.

| Scene | Ground Truth | Rescan | MASC | MASC(fine-tuned) |
|---|---|---|---|---|



*Continued on next page*

Table 5 – *Continued from previous page*

| Scene | Ground Truth | Rescan | MASC | MASC(fine-tuned) |
|---|---|---|---|---|
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |

*Continued on next page*

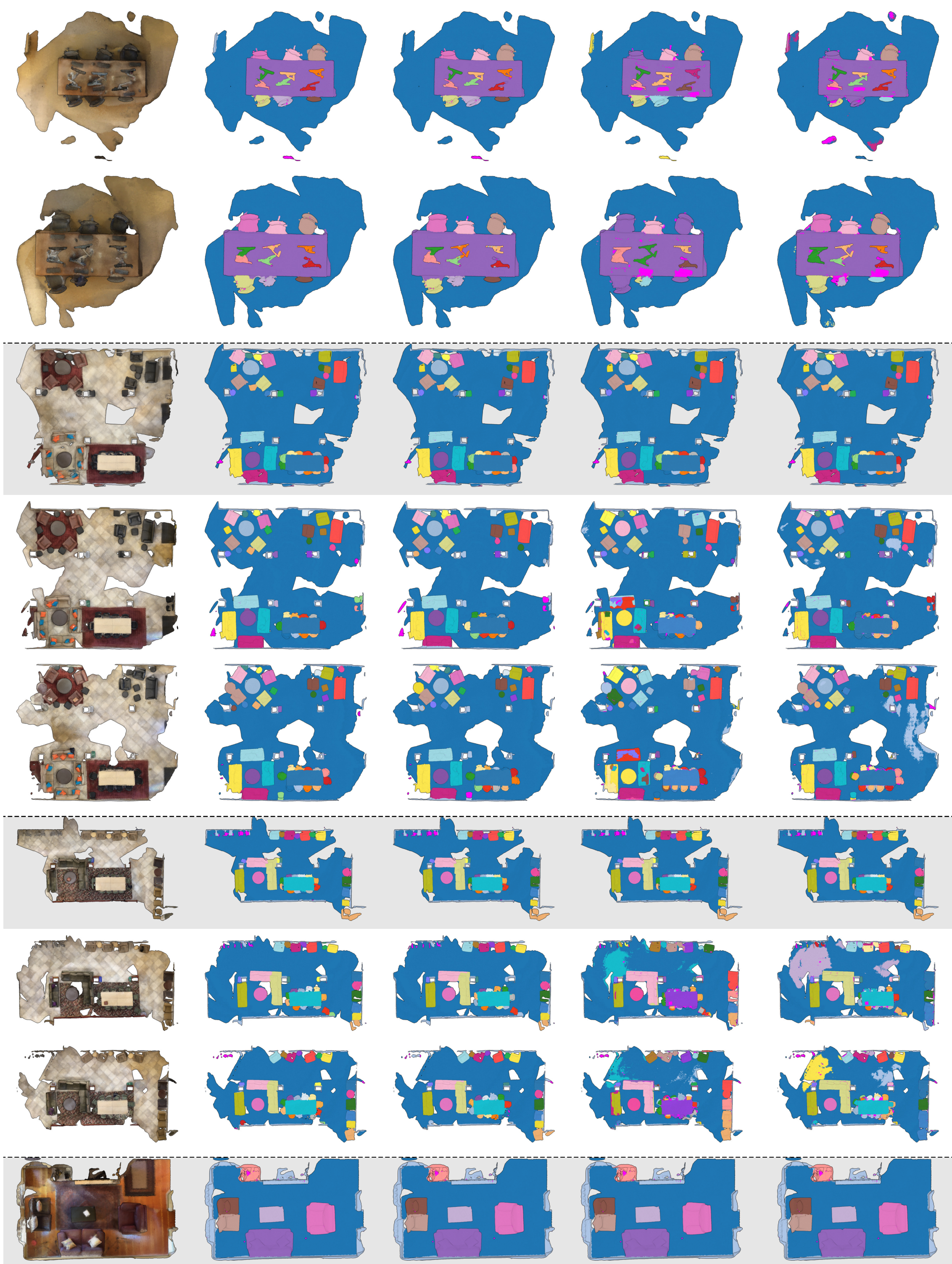
Table 5 – *Continued from previous page*

| Scene | Ground Truth | Rescan | MASC | MASC(fine-tuned) |
|---|---|---|---|---|

*Continued on next page*

Table 5 – *Continued from previous page*

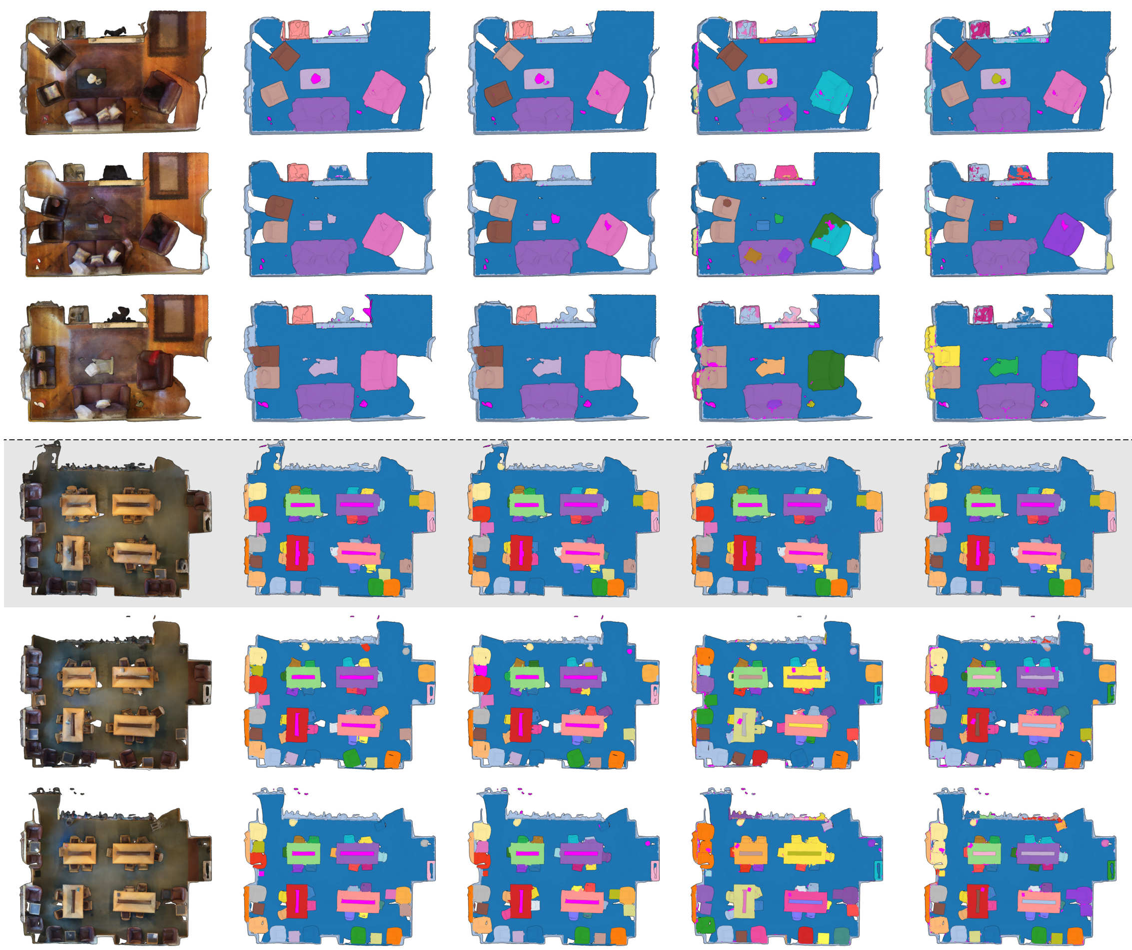
| Scene | Ground Truth | Rescan | MASC | MASC(fine-tuned) |
|---|---|---|---|---|



*Continued on next page*

Table 5 – *Continued from previous page*

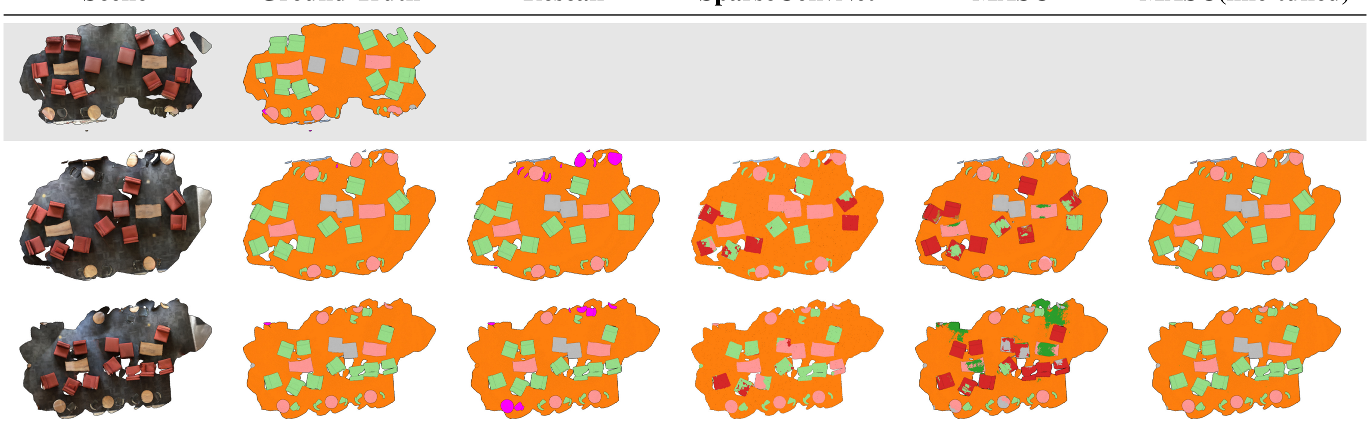
| Scene | Ground Truth | Rescan | MASC | MASC(fine-tuned) |
|---|---|---|---|---|



## A.2. Semantic Label Task

Table 6: *Semantic Label* Task. In the first row for each scene (gray background) we provide reference of the scene $S_0$

| Scene | Ground Truth | Rescan | SparseConvNet | MASC | MASC(fine-tuned) |
|---|---|---|---|---|---|



*Continued on next page*

Table 6 – *Continued from previous page*

| Scene | Ground Truth | Rescan | SparseConvNet | MASC | MASC(fine-tuned) |
|---|---|---|---|---|---|

*Continued on next page*

Table 6 – *Continued from previous page*

| Scene | Ground Truth | Rescan | SparseConvNet | MASC | MASC(fine-tuned) |
|---|---|---|---|---|---|

*Continued on next page*

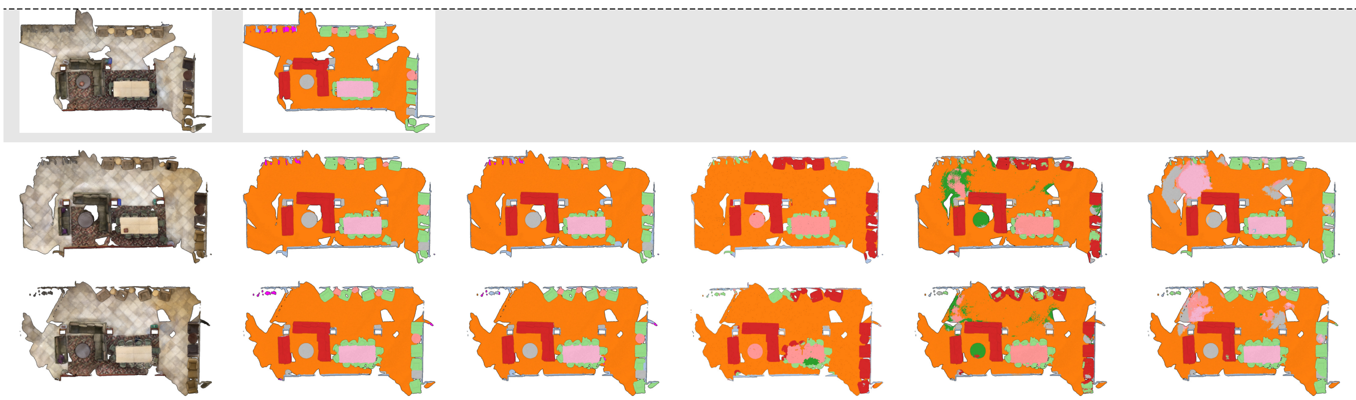
Table 6 – *Continued from previous page*

| Scene | Ground Truth | Rescan | SparseConvNet | MASC | MASC(fine-tuned) |
|---|---|---|---|---|---|

*Continued on next page*

Table 6 – *Continued from previous page*

| Scene | Ground Truth | Rescan | SparseConvNet | MASC | MASC(fine-tuned) |
|---|---|---|---|---|---|



*Continued on next page*

Table 6 – *Continued from previous page*

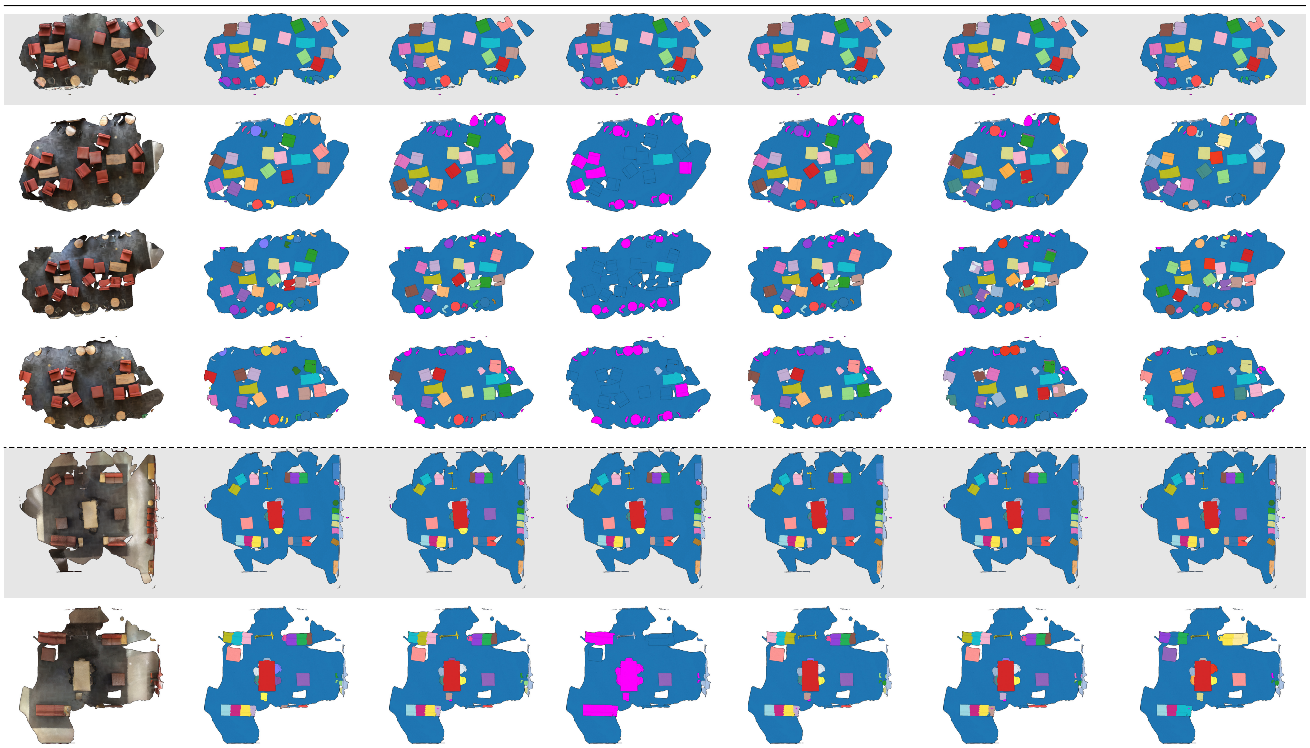
| Scene | Ground Truth | Rescan | SparseConvNet | MASC | MASC(fine-tuned) |
|---|---|---|---|---|---|

# B. Ablation Study Results

## B.1. Semantic Instance Transfer Task

Table 7: *Semantic Instance* and *Instance Transfer Task*. In the first row for each scene (gray background) we visualize ground truth segmentation at time $t_0$. Following rows showcase visualization of instance segmentation estimations. Colors indicate temporal association.

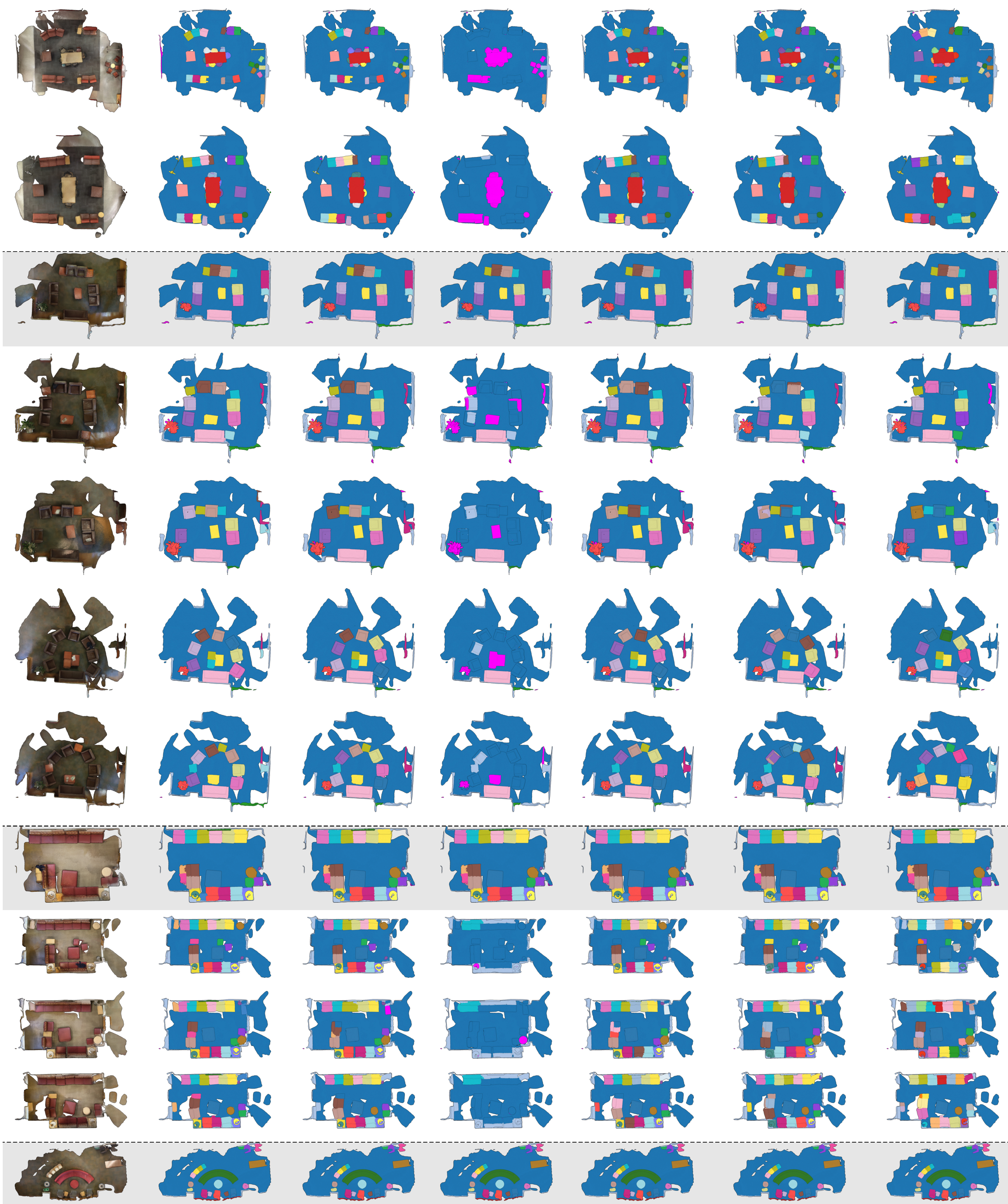
| Scene | Ground Truth | Rescan | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|



*Continued on next page*

Table 7 – *Continued from previous page*

| Scene | Ground Truth | Rescan | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|



*Continued on next page*

Table 7 – *Continued from previous page*

| Scene | Ground Truth | Rescan | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|

*Continued on next page*

Table 7 – *Continued from previous page*

| Scene | Ground Truth | Rescan | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|

*Continued on next page*

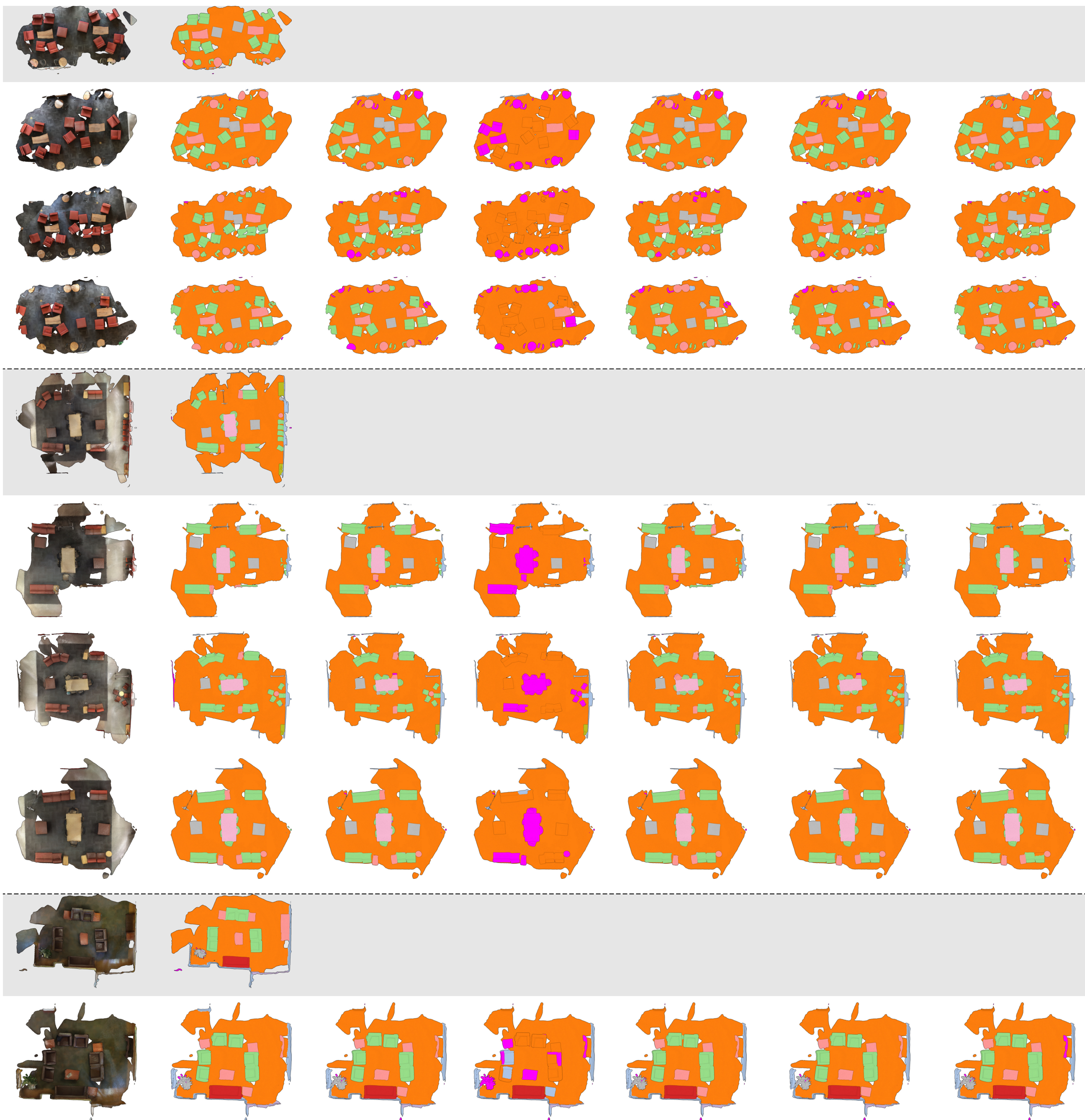
Table 7 – *Continued from previous page*

| Scene | Ground Truth | Rescan | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|

## B.2. Semantic Label Task

Table 8: *Semantic Label* Task. In the first row for each scene (gray background) we provide reference of the scene $S_0$

| Scene | Ground Truth | All Terms | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|



*Continued on next page*

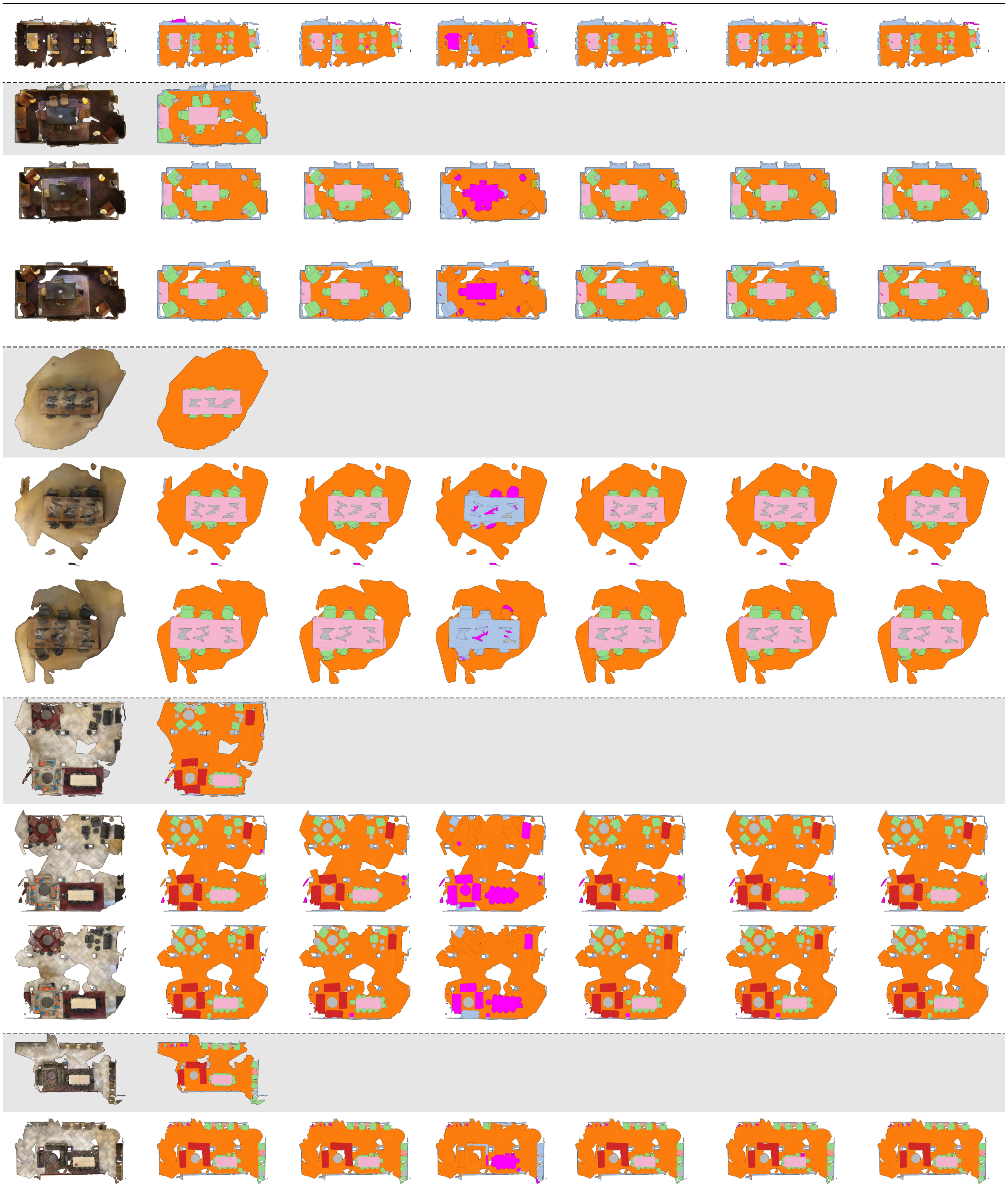
Table 8 – *Continued from previous page*

| Scene | Ground Truth | All Terms | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|

*Continued on next page*

Table 8 – *Continued from previous page*

| Scene | Ground Truth | All Terms | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|



*Continued on next page*

Table 8 – *Continued from previous page*

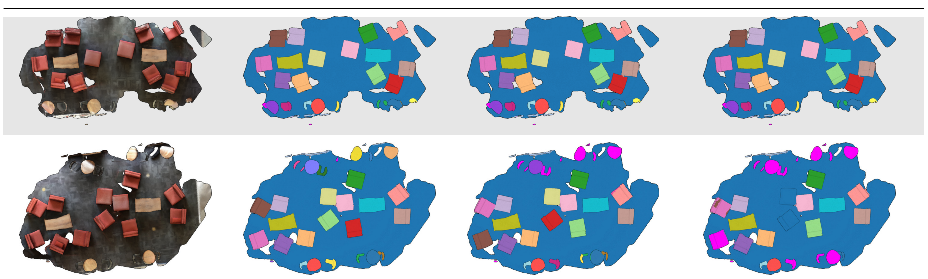
| Scene | Ground Truth | All Terms | No Coverage | No Geometry | No Intersection | No Hysteresis |
|---|---|---|---|---|---|---|

## C. Limited Movement Study Results

### C.1. Semantic Instance Transfer Task

Table 9: *Semantic Instance* and *Instance Transfer Task*. In the first row for each scene (gray background) we visualize ground truth segmentation at time $t_0$. Following rows showcase visualization of instance segmentation estimations. Colors indicate temporal association.

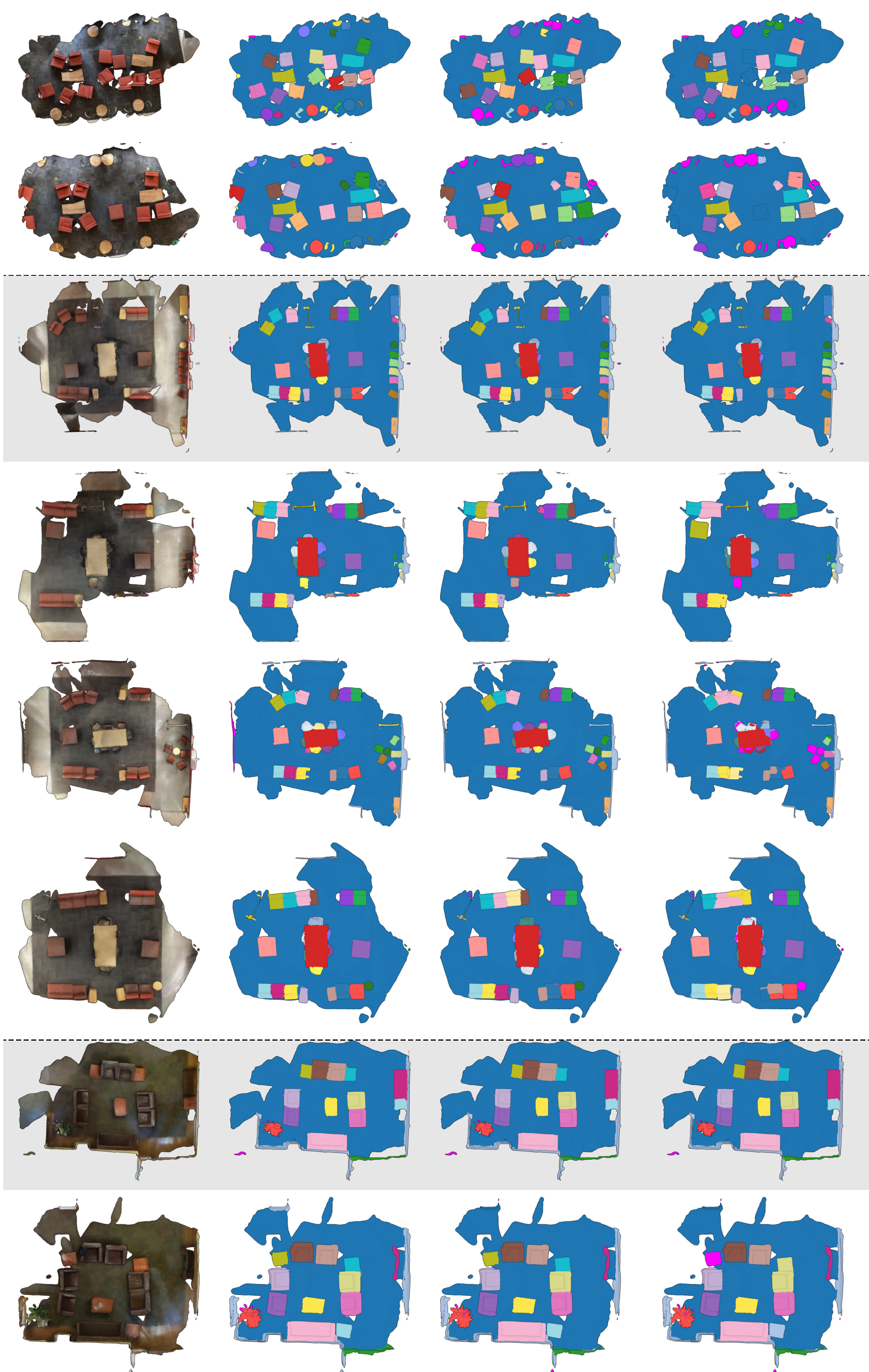
| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|



*Continued on next page*

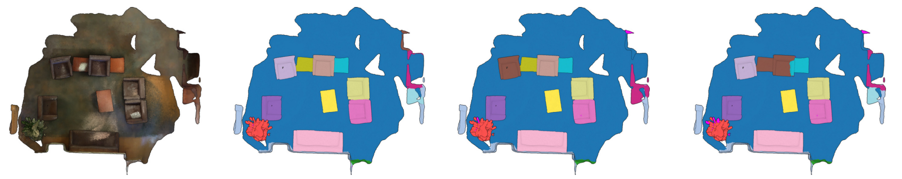
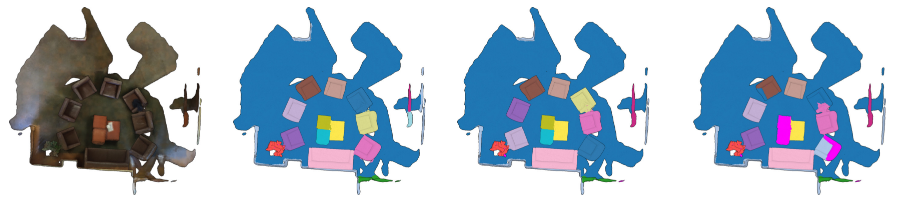
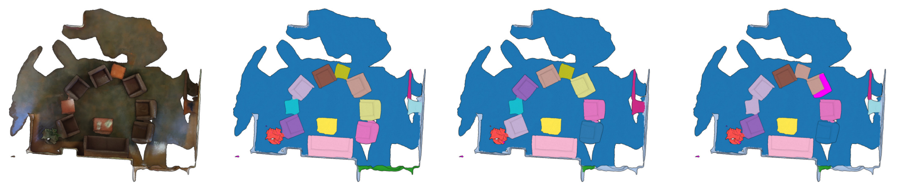
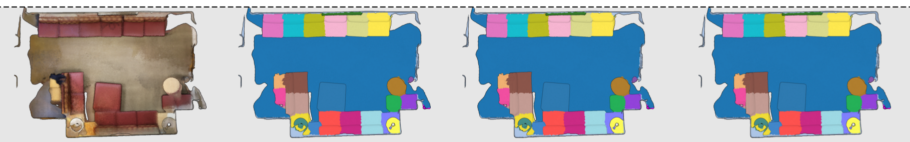
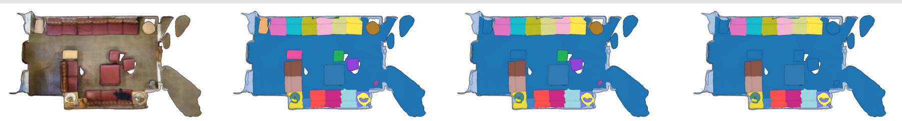
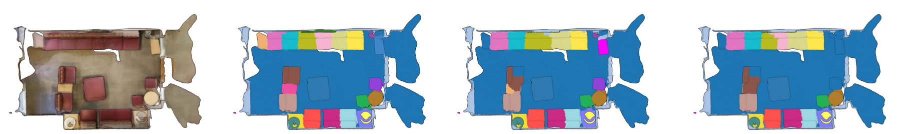
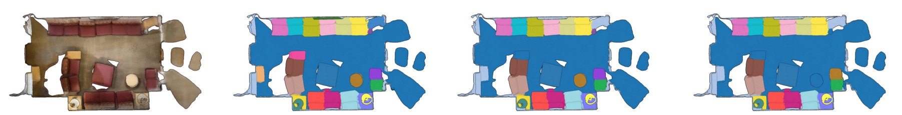
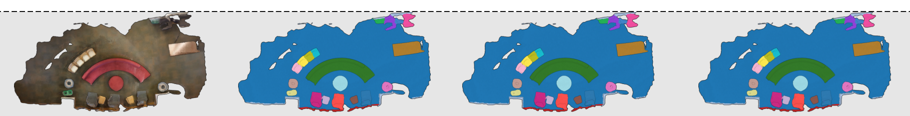
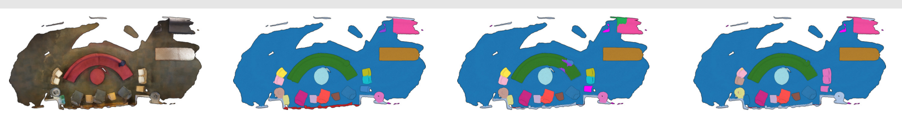
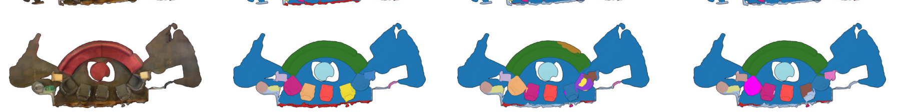
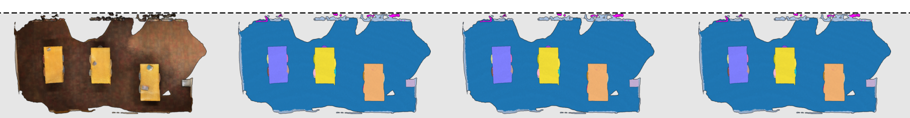
Table 9 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|



*Continued on next page*

Table 9 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |
|  | | | |

*Continued on next page*

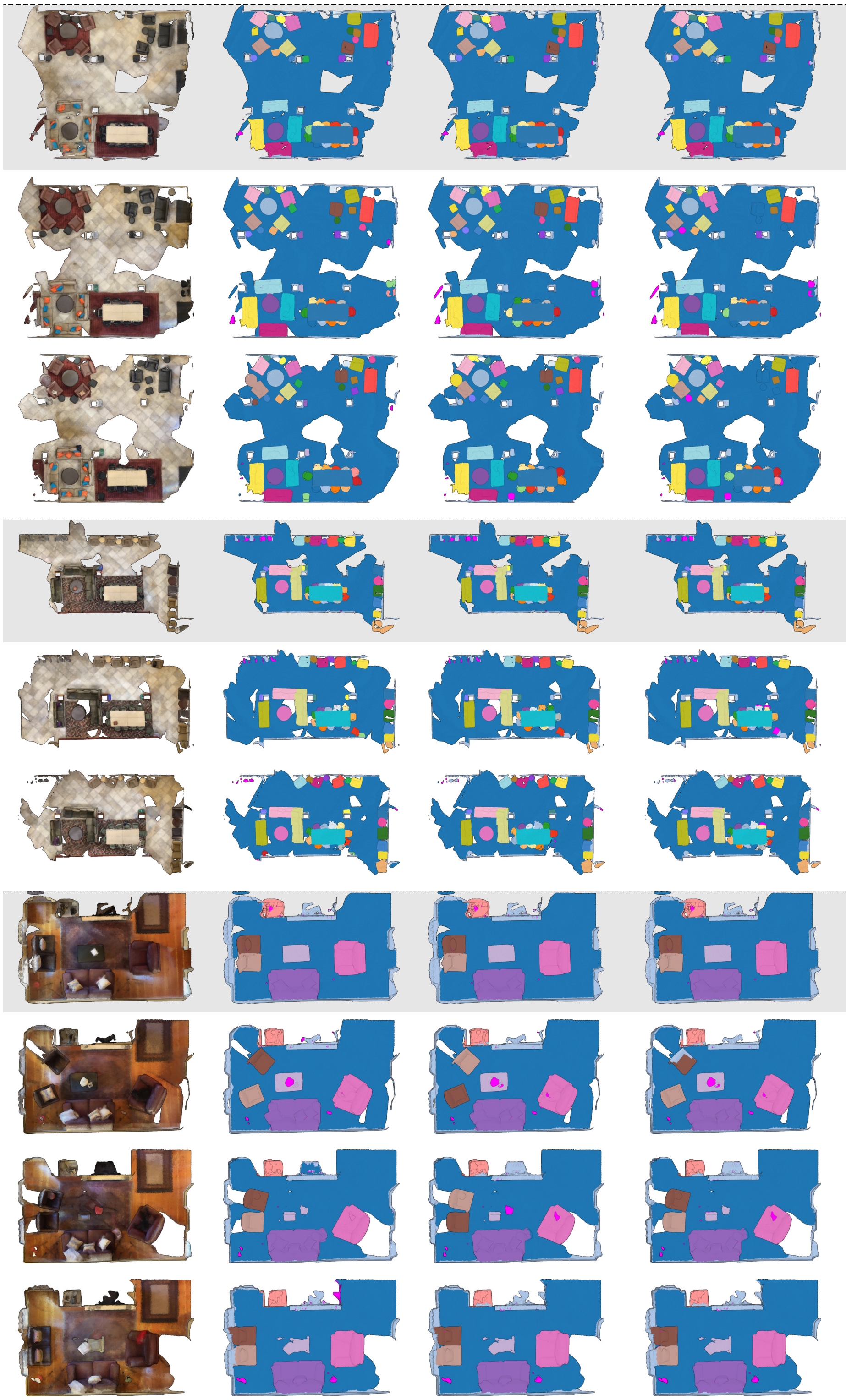
Table 9 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|

*Continued on next page*

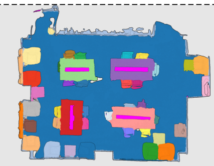
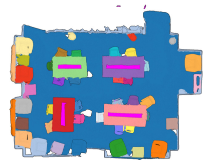
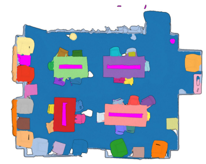
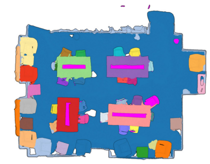
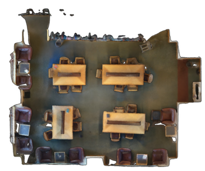
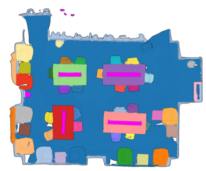
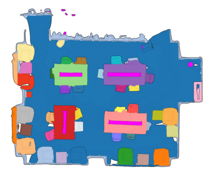
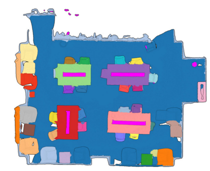
Table 9 – *Continued from previous page*

| **Scene** | **Ground Truth** | **Full Movement** | **Limited Movement** |
|---|---|---|---|



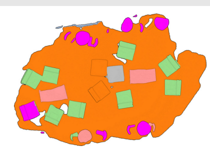
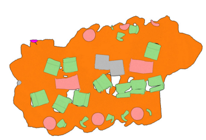
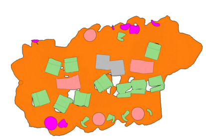
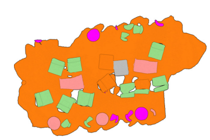
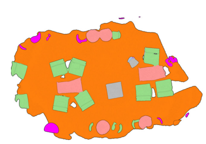
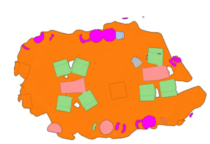
*Continued on next page*

Table 9 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

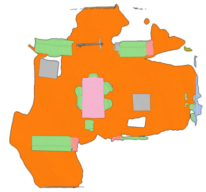
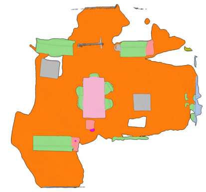
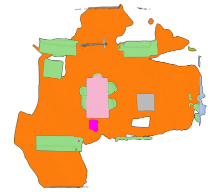
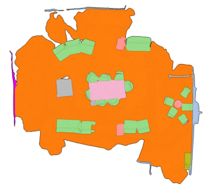
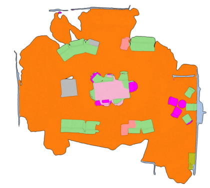
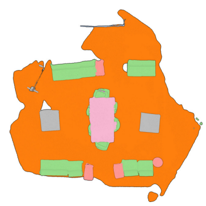
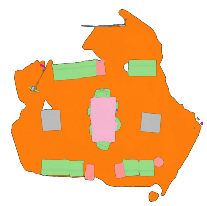
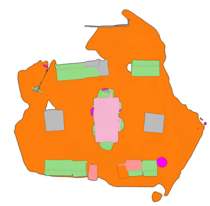
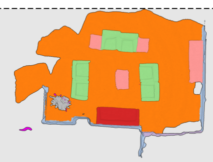
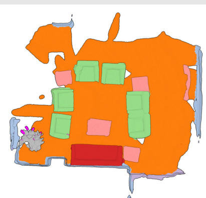
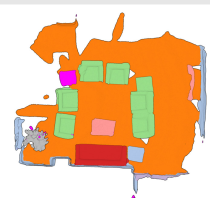
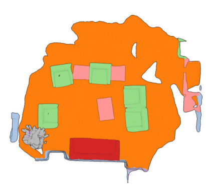
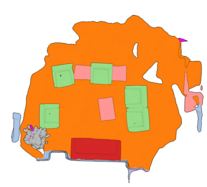
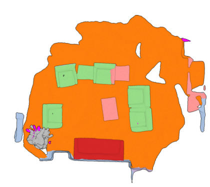
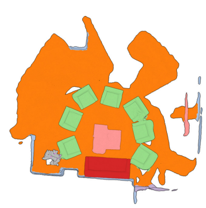
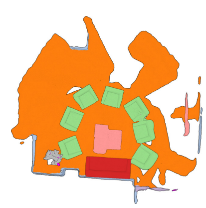
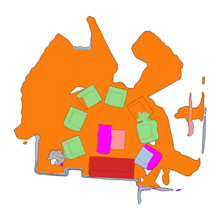
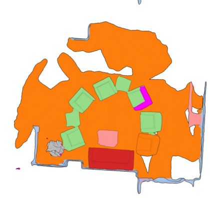
### C.2. Semantic Label Task

Table 10: *Semantic Label* Task. In the first row for each scene (gray background) we provide reference of the scene $S_0$

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|
|  |  | | |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  | | |

*Continued on next page*

Table 10 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  | | |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

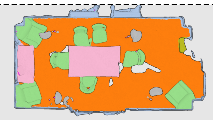
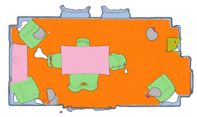
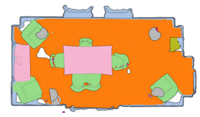
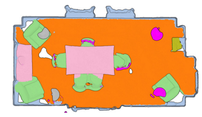
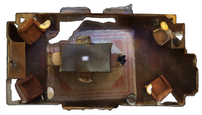
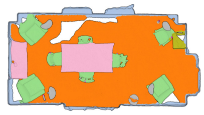
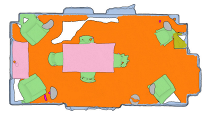
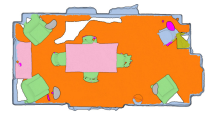
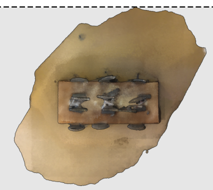
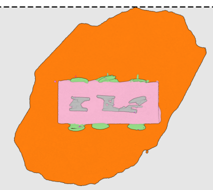
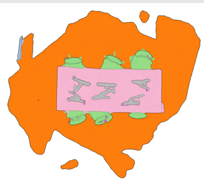
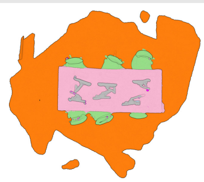
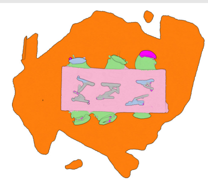
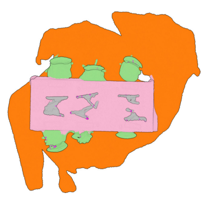
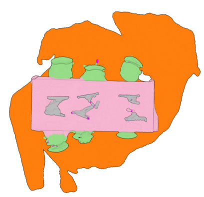
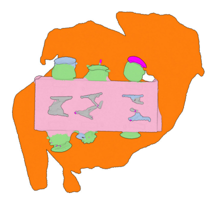
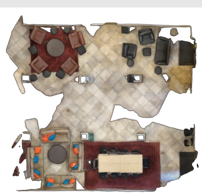
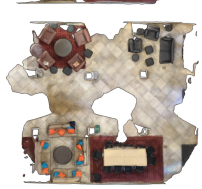
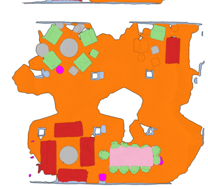
*Continued on next page*

Table 10 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|

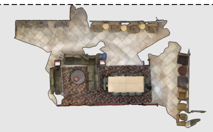
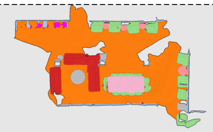
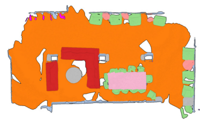
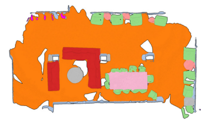
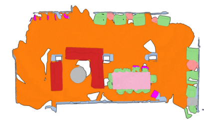
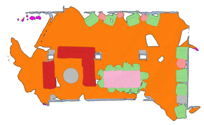
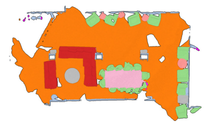
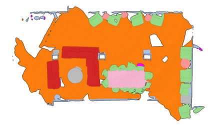
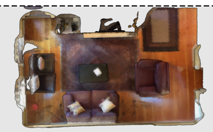
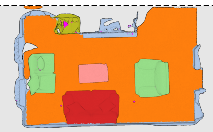
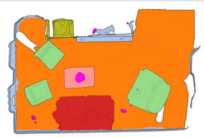
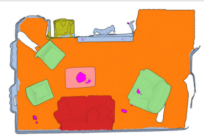
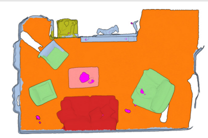
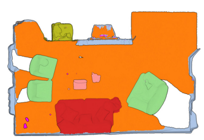
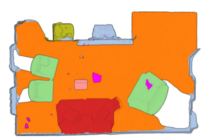
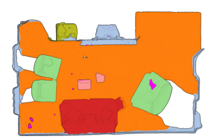
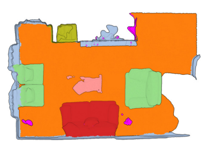
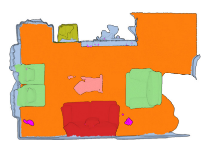
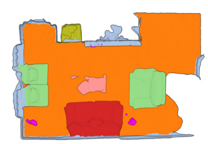
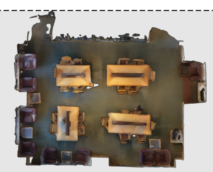
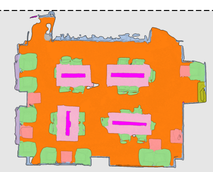
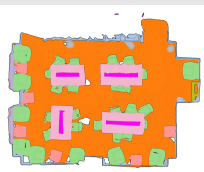
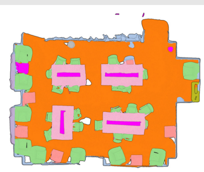
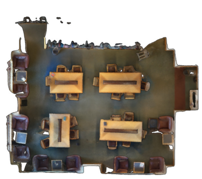
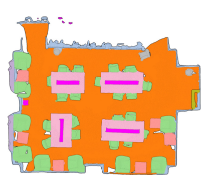
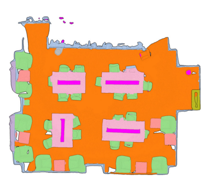
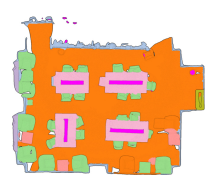
*Continued on next page*

Table 10 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|
|  |  | | |
|  |  |  |  |
|  |  |  |  |
|  |  | | |
|  |  |  |  |
|  |  |  |  |
|  |  | | |
|  |  |  |  |
|  |  |  |  |

*Continued on next page*

Table 10 – *Continued from previous page*

| Scene | Ground Truth | Full Movement | Limited Movement |
|---|---|---|---|
|  |  | | |
|  |  |  |  |
|  |  |  |  |
|  |  | | |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  | | |
|  |  |  |  |
|  |  |  |  |

## References


[1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[2] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2

[3] M. Halber and T. Funkhouser. Fine-to-coarse global registration of rgb-d scans. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1

[4] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, M.-K. Tran, L.-F. Yu, and S.-K. Yeung. Scenenn: A scene meshes dataset with annotations. In *International Conference on 3D Vision (3DV)*, 2016. 2