# Rescan: Inductive Instance Segmentation for Indoor RGBD Scans

Maciej Halber Yifei Shi Kai Xu Thomas Funkhouser
Princeton University

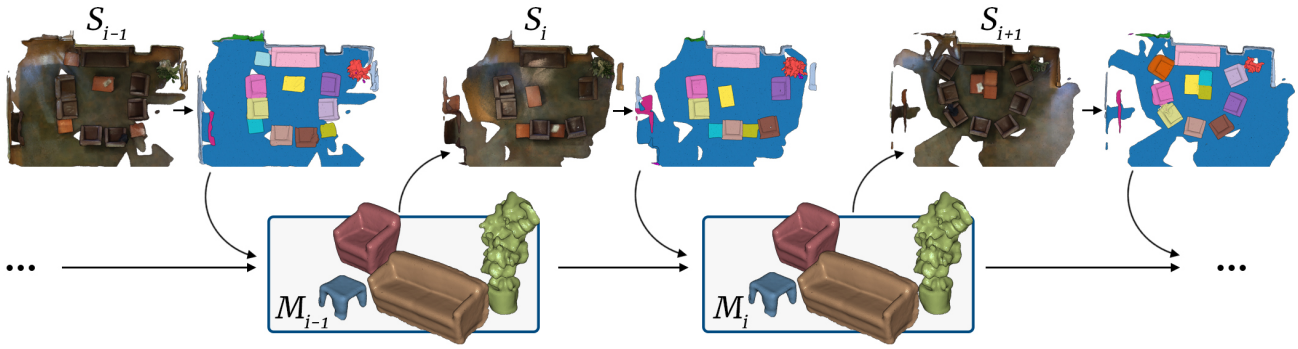



Figure 1: The proposed method estimates a persistent, temporally-aware scene model $M_i$ from a series of scene observations $S_i$, captured at sparse time intervals. $M_{i-1}$ is used to estimate an arrangement of objects in each novel observation $S_i$. The estimated arrangement is used to estimate the instance segmentation of $S_i$, which is then used to update the model $M_i$.

## Abstract

*In depth-sensing applications ranging from home robotics to AR/VR, it will be common to acquire 3D scans of interior spaces repeatedly at sparse time intervals (e.g., as part of regular daily use). We propose an algorithm that analyzes these "rescans" to infer a temporal model of a scene with semantic instance information. Our algorithm operates inductively by using the temporal model resulting from past observations to infer an instance segmentation of a new scan, which is then used to update the temporal model. The model contains object instance associations across time and thus can be used to track individual objects, even though there are only sparse observations. During experiments with a new benchmark for the new task, our algorithm outperforms alternate approaches based on state-of-the-art networks for semantic instance segmentation.*

## 1. Introduction

With the proliferation of RGBD cameras, 3D data is now more widely available than ever before [10, 25, 8]. As depth capturing devices become smaller and more affordable, and as they operate in everyday applications (AR/VR, home robotics, autonomous navigation, etc.), it is plausible to expect that 3D scans of most environments will be acquired on a daily basis. We can expect that 3D reconstructions of many spaces, visited at different times and captured from different viewpoints, will be available in the future, just like photographs are today.

In this paper, we investigate how repeated, infrequent scans captured with handheld RGBD cameras can be used to build a spatio-temporal model of an interior environment, complete with object instance semantics and associations across time. The challenges are that: 1) each RGBD scan captures the environment from different viewpoints, possibly with noisy data; and 2) scans separated by long time intervals (once per day, every Tuesday, etc.) can have large differences due to object motion, entry, or removal. Thus simple algorithms that perform object detection individually for each scan and/or simply cluster object detections and poses in space-time will not solve the problem. Moreover, since large training sets are not available for this task, it is not practical to train a neural network to solve it.

We propose an inductive algorithm that infers information about new RGBD capture of a scene $S_i$ from a temporal model $M_{i-1}$ obtained from previous observations of $S$ (fig. 1). The input to the algorithm is the model $M_{i-1}$, representing all previous scans and a novel scene scan $S_i$. The output is an updated model $M_i$ that describes the set of objects $\mathcal{O}$ appearing in the scene and an arrangement $\mathcal{A}$ of those objects at each time step, including the most recent. At every iteration, our algorithm optimizes for the arrangement $A_i$ of objects in $S_i$, and then uses $A_i$ to infer the semantic instance segmentation of $S_i$. Segmentation of $S_i$ is then used to update object set $\mathcal{O}$ (see fig. 2).

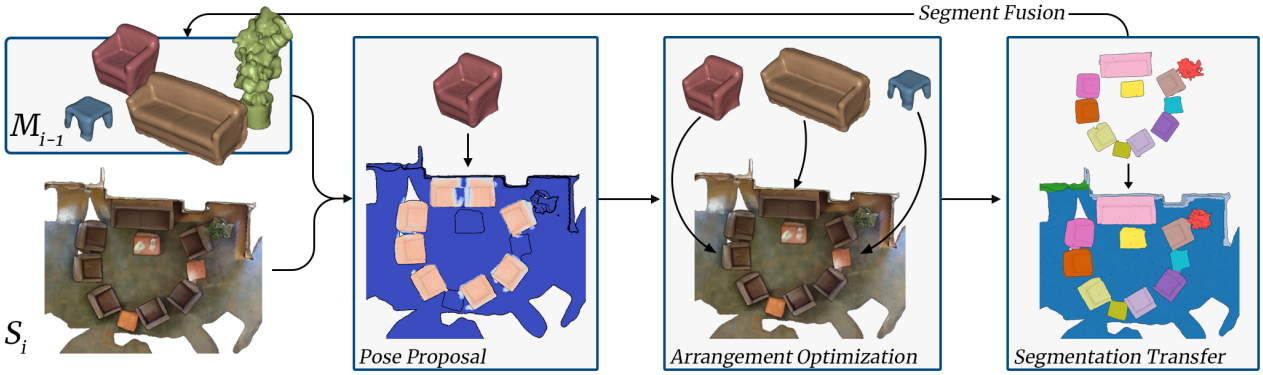



Figure 2: A single inductive step of the proposed method. Given a novel scene observation $S_i$ and a model from the past $M_{i-1}$, our goal is to create an updated model $M_i$. We first perform *Pose Proposal*, where we search for a set of potential locations for each object in $M_{i-1}$ . Then, we perform *Arrangement Optimization*, where we search for the selection and arrangement of objects to minimize an objective function. Then, we perform *Segmentation Transfer*, in which $S_i$ is annotated with semantic instance labels from $M_{i-1}$. Finally, geometry from segments in $S_i$ is fused with $M_{i-1}$ to create an updated model $M_i$.

To evaluate our algorithm we present a novel benchmark dataset that contains temporally consistent ground-truth semantic instance labels, describing object associations across time within each scene. Experiments with this benchmark suggest that our proposed optimization strategy is superior to alternative approaches based on deep learning for semantic and instance segmentation tasks.

Overall, the contributions of the paper are three-fold:

- A system for building a spatio-temporal model for an indoor environment from infrequent scans acquired with hand-held RGBD cameras,
- An inductive algorithm that jointly infers the shapes, placements, and associations of objects from infrequent RGBD scans by utilizing data from past scans,
- A benchmark dataset with rescans of 13 scenes acquired at 45 time-steps in total, along with ground-truth annotations for object instances and associations across time.

## 2. Related Work

Most work in computer vision on RGBD scanning of dynamic scenes has focused on tracking [43] and reconstruction [36]. For example, Newcombe et al. [36] showcases a system where multiple observations of a deforming object are fused into a single consistent reconstruction. Yan et al. [48] scan moving articulated shapes by tracking parts as they are deformed over time. These methods differ from ours as they require observation of motions as they occur.

For sparse temporal observations, early work in robotics focuses on the analysis of 2D maps created from 1D laser range sensors [3, 5, 19]. For example, Biswas [5] used 1D laser data to detect objects within a scene and associate them across time. However, their method relies upon 2D algorithms and assumes that object instances cannot overlap across time, which makes it inapplicable in our setting. More recently, image based techniques for sparse observations were proposed — Shin [42] extends SfM to also predict poses of moving objects.

Other work has aimed at life-long scene understanding using data captured with actively controlled sensors [15, 29, 39, 49]. For example, several algorithms proposed in the STRANDS project [23] process the scenes observed from a repeated set of views [2, 6, 41]. Others focus on controlling camera trajectories to acquire the best views for object modeling [13, 15] and/or change detection [1]. These problems are different than ours, as we focus on analyzing previously acquired RGBD data captured without a specifically tailored robotic platform and active control.

Some work in computer vision has focused on change detection and segmentation of dynamic objects in RGBD scans [16, 31, 47]. For example, Fehr et al. [16] showcases a system for using multiple scene observations to classify surface elements as dynamic or static. Wang et al.[46] detect moving objects so that they can be removed from a SLAM optimization. Lee et al. [31] propose a probabilistic model to isolate temporally varying surface patches to improve camera localization. While operating on RGBD captures from handheld devices, these methods do not produce instance-level semantic segmentations, nor do they generate associations between objects across time.

More recent work has focused on automatic clustering of 3D points into clusters across space and time [17, 24]. For example, Herbst et al. [24] jointly segments multiple RGBD scans with a joint MRF formulation. Finman et al. [17] detects clusters of points from pairwise scene differencing and associates new detections with previous observations. Although similar in spirit to our formulation, these methods

operate only on clusters of points, without semantics, and thus are not suited for applications that require semantic understanding of how objects move across space-time.

Finally, many projects have considered temporal modeling of environments in specific application domains. For example, several systems in civil engineering track changes to a Building Information Model (BIM) by alignment to 3D scans acquired at sparse temporal intervals [20, 26, 37, 45]. They generally start with a specific building design model [22], construction schedule [44], and/or object-level CAD models [7], and thus are not as general as our approach. The Scene Chronology project [35] and others [34, 40] build temporal models of cities from image collections – however, they do not recover a full 3D model with temporal associations of object instances as we do.

# 3. Algorithm

## 3.1. Scene Representation

Our system represents a scene at time $t_i$ with a temporal model $M_i$ comprising a tuple $\{\mathcal{O}, \mathcal{A}\}$, where $\mathcal{O} = \{o_0, \ldots, o_n\}$ is a list of $n$ object instances that have appeared within this or any prior observation $S_j$ for $j \in [0, i]$, and $\mathcal{A} = \{A_0, \ldots, A_i\}$ is a list of object arrangements estimated for each observation $S_j$. Each object instance $o_k$ is represented by $\{u_k, G_k, c_k\}$, where $u_k$ is unique instance id, $G_k$ is the object's geometry, and $c_k$ is the semantic class. Each arrangement $A_i$ is a list of poses $\{a_i^0, \ldots, a_i^m\}$, where $a_i^j = \{u_j, \mathbf{T}_j, s_j\}$. $u_j$ is the unique id of $j$-th object and function $\Omega(u_j)$ returns index $k$ to $\mathcal{O}$. $\mathbf{T}_j$ is a transformation that moves geometry $G_k$ into correct location within the scene $S_i$. Lastly $s_j$ is a matching score quantifying how well $\mathbf{T}_j G_k$ matches the geometry of $S_i$.

## 3.2. Overview

Our algorithm updates the temporal model in an inductive fashion. Given the previous model $M_{i-1}$ and a new scan $S_i$, we predict a new model $M_i$ (see fig. 2) by executing four consecutive steps. The first proposes potential poses for objects in $\mathcal{O}$ (sec. 3.3). The second performs a combinatorial optimization to find the arrangement $A_i$ that maximizes a new objective function jointly accounting for geometric fit and temporal coherence (sec. 3.4). The third step uses $\mathcal{O}$ and $A_i$ to infer an instance-level semantic segmentation of $S_i$. The fourth step updates the geometry $G_k$ of each object $\in A_i$ by aggregating its respective segment from $S_i$. The following four subsections offer the details on how each of these steps is implemented.

## 3.3. Object Pose Proposal

The first step of our pipeline is to find a set of potential placements for each object $o_k \in \mathcal{O}$, creating a search space for the Arrangement Optimization stage (sec. 3.4). Formally, the input to this stage is a set of objects $\mathcal{O}$ and a scan $S_i$. The output is a set $\mathcal{P}$ of scored pose lists $P_k = \{p_k^0, \ldots, p_k^x\}$ for each object $o_k$. A scored pose $p_k^l$ is a tuple $\{\mathbf{T}_k^l, s_k^l\}$, where $\mathbf{T}_k^l$ is the proposed rigid-body transformation and $s_k^l$ is a geometric matching score describing how well pose $\mathbf{T}_k^l$ aligns $G_k$ to the geometry of $S_i$.

Finding transformations that align surfaces $A$ and $B$ is a longstanding problem in computer graphics and vision [38]. In our setting, we wish to find a set of poses for the surface $A$ with good alignment with surface $B$, where $A = o_k$ and $B = S_i$. Prior work usually attempts to solve similar problems by employing feature-based methods. Such methods sub-sample the two surfaces to obtain a set of meaningful keypoints and then match them to produce a plausible pose (e.g., using Point-Pair Feature matching[12]). However, as it has been noted in other domains, keypoints may limit the amount of information a method considers, with dense matching methods leading to less failures [14].

Following this intuition, we propose a dense matching procedure, where we slide each of the objects $o_k$ across the scene, perform an ICP optimization at each of the discrete locations and compute a matching score based on the traditional point-to-plane distance metric [32].

This approach might seem counter-intuitive, as a naive implementation of such grid-search would lead to a prohibitive run-time performance. We find however that such an approach can be made acceptably fast while leading to much better recovery of correct poses. To speed-up the run-time performance of our method we make use of the multi-resolution approach. We compute a four-level hierarchy for the input point cloud (the geometries $G_k$), with minimum distance between any two points at a level equal to $\{0.01m, 0.02m, 0.04m, 0.08m\}$ respectively. To compute this representation we follow an algorithm described in [9]. Multi-resolution representation allows us to perform the dense search only on the coarsest level of the hierarchy, and return a subset of poses with sufficiently high scores to be verified at higher levels, leading to significant performance gains. Additionally, we make a simplifying, but reasonable assumption that objects in our scenes move on the ground plane and rotate around the gravity direction.

With this approach we are able to produce a set $\mathcal{P}$ of pose lists $P_k$ for each object $o_k$ in $\mathcal{O}$. The advantage of this dense grid-search method is that it produces sets of poses that contain most of the true candidate locations, even if the local geometry of $S_i$ might be different from $G_k$ due to reconstruction errors. We showcase the comparison to keypoint based methods [12, 4] in figure 3.

## 3.4. Arrangement Optimization

In the second step our algorithm selects a subset of poses from the previous step to form an object arrange-

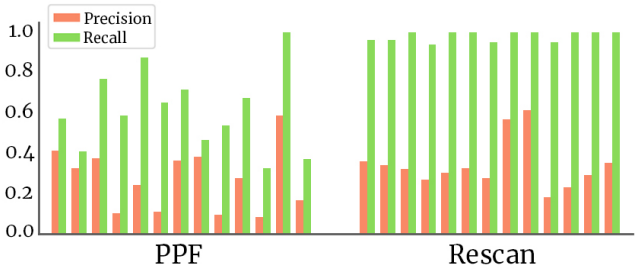



Figure 3: Comparison of the precision/recall scores obtained for all scenes in our database, comparing PPF matching [4] to our method. In our experiments a pose of an object $o_k$ is considered a true positive if the distance between object centers is less than $0.2m$ and object's classes agree.

ment. The input is a set of objects $\mathcal{O}$, a set of pose lists $\mathcal{P} = \{P_0, \ldots, P_k\}$ for each object $o_k$, and the scan $S_i$. The output is an arrangement $A_i$ that describes a global configuration of objects which maximizes the objective.

This problem statement leads to a discrete, combinatorial optimization. First reason for choosing this approach is that the number of objects within the scene $S_i$ is not known a priori. A combinatorial approach allows us to propose arrangements $A_i$ of variable lengths, that will adapt to the contents of $S_i$. A second reason is that finding the optimum requires global optimization – the placement of one object can greatly affect the placement of another. Additionally, deep learning is hard to apply in this instance due to the lack of the training data, as well as the non-linearity of the proposed objective function.

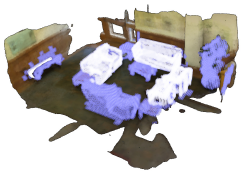
#### 3.4.1 Objective Function

To quantify the quality of the candidate arrangement $A'_i$ we use the objective function that is a linear combination of the following four terms:

$$\begin{aligned} O(S_i, A'_i, \mathcal{A}) = w_c O_c(S_i, A'_i) & \quad \text{Coverage Term} \\ + w_g O_g(S_i, A'_i) & \quad \text{Geometry Term} \\ + w_i O_r(A'_i) & \quad \text{Intersection Term} \\ + w_h O_h(A'_i, \mathcal{A}) & \quad \text{Hysteresis Term} \end{aligned}$$

Each term $O_x$ produces a scalar value $\in [0, 1]$ that describes the quality of $A'_i$ w.r.t. that specific term. We use grid search to find good values for the weights $\mathbf{w} = \{2.0, 0.3, 1.0, 1.8\}$, which express the relative importance of each term.

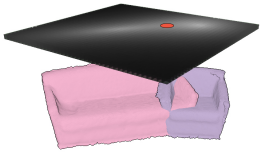
**The Coverage term** measures the percentage of the scene that is covered by objects in $A'_i$. The intuition behind this term is that every part of the scene should ideally be explained by some object in $A'_i$.



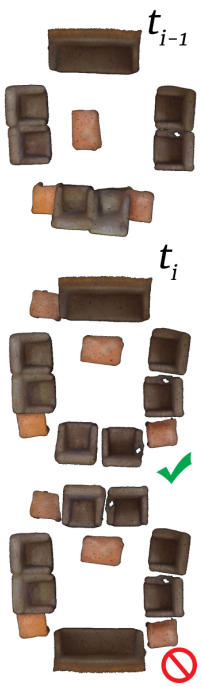
$O_c(S_i, A'_i)$ takes as input a scene $S_i$ and the candidate arrangement $A'_i$. To compute $O_c(S_i, A'_i)$ we voxelize both the scene $S_i$ and the objects in $A'_i$, resulting in two 3D grids $V_S$ and $V_A$. The $O_c(S_i, A'_i)$ is calculated as the number of cells that are equal in both grids, over the number of cells in $V_S$ - $O_c(S_i, A'_i) = \frac{|V_s(j) \wedge V_A(j)|}{|V_s(j)|}$. For this formula to be accurate we need to ensure however that we only voxelize the dynamic parts of the scene $S_i$. As such we deactivate any cells in $V_S$ that belong to the static parts of the scene, like walls and floor, which can easily be detected with a method like RANSAC [18]. The inset figure above showcases a visualization of both grids $V_S$ (blue cells) and $V_A$ (white cells). As seen there, the $V_S$ covers the non-static parts of the scene only, leading to $O_c$ being a good estimate of the coverage.

**The Geometry term** is a measure of the geometrical agreement between the scene $S_i$ and objects in the candidate arrangement $A'_i$. We include this term to guide the objective function to select objects that best match the geometry of the scene at a specific location. This value is simply computed as an average of scores $s^l_k$ from the procedure described in section 3.3. $O_g(S_i, A'_i) = \frac{\sum_k g(a^j_i)}{|A'_i|}$, where $g(a^j_i)$ returns the geometrical score fit for placement of object $o_j$.

**The Intersection term** aims to estimate how much a pair of objects in the arrangement $A'_i$ interpenetrate. Intuitively, such interpenetration would mean that two objects occupy the same physical location, which implies an impossible configuration. In our approach, we compute a coarse approximation of this term. First, we compute a covariance matrix $\mathbf{\Sigma}_k$ of each $G_k$. Covariances for each object allow us to compute a symmetric Mahalanobis distance $SD_M$ between objects to approximately quantify how close they are to each other. $SD_M(O_r, o_j) = 0.5(D_M(m_{ij}, \mathbf{T}_i c_i, \Sigma_i) + D_M(m_{ij}, \mathbf{T}_j c_j, \Sigma_j))$, where $\mathbf{T}_i c_i$,$\mathbf{T}_j c_j$ are transformed centroids of $G_i$, $G_k$, the midpoint between them is $m_{ij}$, and function $D_M$ is the Mahalanobis distance. With $SD_M$ computed for all pairs of objects $o_k$, the value $O_r(A'_i)$ is $1 - ||\{\exp(\frac{-SD^2_M(o_0,o_1)}{2\sigma^2}), \ldots, \exp(\frac{-SD^2_M(o_{n-1},o_n)}{2\sigma^2})\}||_\infty$. The rationale behind the use of the infinity norm is to generate a high penalty if just a single pair of objects exhibits a low score interpenetration. The inset figure above showcases a visualization of $SD_M$ for two intersecting objects. The point at which we evaluate the $SD_M$ is marked with red, showcasing high values in regions where either or both objects are present, and low values in the free space. It is also clear that the value of $SD_M$ would be higher if the objects interpenetrated even more.



**The Hysteresis term** informs how well the current arrangement estimate $A'_i$ resembles a previously observed arrangements from the set $\mathcal{A}$. In addition it expresses our preference for a minimal relative motion. Each object in $A'_i$ is assigned a score, with the value based on whether $u_k$ is a novel instance, or has been observed in the past. In the former case, we assign a novel object constant score $h = 0.4$ (found manually). In the latter, the score is $h + (1 - h)exp(\frac{-||T(c_k,i)-T(c_k,j)||_2}{2\sigma^2})$. $T(c_l, j)$ is a function that applies the appropriate transformation to centroid $c_l$ at time $t_j$. As a result, novel objects will be always preferred, unless they have undergone a significant transformation. In such a case, we would like $O_h$ to express that novel object appearances have similar probability. The value of $O_h(A_i, \mathcal{A})$ is computed as an average of the above scores. The inset figure above illustrates an arrangement at $t_{i-1}$ and two possible arrangement estimates at $t_i$. The form of $O_h(A'_i, \mathcal{A})$ encourages the selection of middle arrangement as it does not contain significant motion the sofa and chairs.




#### 3.4.2 Optimization

To find arrangement $A_i = \arg\max_{A'_i} O(S_i, A'_i, \mathcal{A})$, we employ a combination of greedy initialization and simulated annealing. We begin by greedily selecting an object $o_k$ at a pose $p^l_k$ which improves objective the most. This process of addition is continued until the objective function starts decreasing. After this stage, we perform simulated annealing optimization. We run the simulated annealing for 25k iterations, using a linear cooling schedule with a random restarts ($0.5\%$ probability to return to the best scoring state). To explore the search space we use the following actions with a randomly selected object $o_k$:

- **Add Object** - We add $o_k$ at a random pose $p^l_k$ to $A'_i$.
- **Remove Object** - We remove $o_k$ from $A'_i$.
- **Move Object** - We select $o_k$ from $A'_i$ and assign it new pose $p^m_k$.
- **Swap Objects** - We swap the location of $o_k$ and $o_l$, another randomly selected object of the same semantic class.

### 3.5. Segmentation Transfer

The third step of the algorithm transfers the semantic and instance labels from $A_i$ to scan $S_i$. The estimated arrangement from the previous step can be used to perform segmentation transfer, as we have semantic class $c_k$ and instance id $u_k$ associated with each object in $\mathcal{O}$. Using the estimated pose $p^l_k$ for each of the objects $o_k$ in $A_i$, we transform its geometry $G_k$ to align with $S_i$. We then perform a nearest neighbor lookup (with a maximum threshold $d = 5cm$ to account for outliers) and use the associations to copy both the instance and semantic labels from objects in $A_i$ to $S_i$. Since there is no guarantee that all points in $S_i$ will have a neighbor within the threshold $d$, we follow-up the lookup with label smoothing based on multi-label graph-cut [11].

### 3.6. Geometry Fusion

The final step of the algorithm is to update the object geometries $G_k$ for objects in $\mathcal{O}$. To do so for each object $o_k \in A_i$, we extract the sub point clouds from $S_i$ that were assigned instance label $u_k$ in the previous step, and then we concatenate them with $G_k$ to generate new point cloud $G'_k$. In the idealized case, the two surfaces would be identical, as they represent the same object. However, due to partial observation, reconstruction, and alignment errors, we cannot expect that in practice. As such, we solve for a mean surface $\tilde{G}_k$ that minimizes the distance to all points in the $G'_k$, using Poisson Surface Reconstruction [27]. After this process, we uniformly sample points on the resulting surface $\tilde{G}_k$ to get a new estimate of $G_k$ that will be used for matching when a new scene $S_{i+1}$ needs to be processed.

## 4. Evaluation

Evaluation of the proposed algorithm is not straightforward, as there is little to no prior work directly addressing instance segmentation transfer between 3D scans.

**Dataset:** To evaluate the proposed approach, we have created a dataset of temporally varying scenes. Our dataset contains 13 distinct scenes, with total of 45 separate reconstructions. Each scene contains between 3 to 5 scans, where objects within each catpure were moved to simulate changes occuring across long time periods. Along with the captured data, we also provide manually-curated semantic category and instance labels for every object in every scene. The instance labels are stable across time, providing associations between object instances in different scans, which we can use to evaluate our algorithms. Additionally, we provide permutations of instance assignments for each scene to account for cases where objects' motion is ambiguous and multiple arrangements can be considered correct. More details about the dataset are included in the supplemental material.

**Metrics:** We evaluate our approach using three metrics. The first is the *Semantic Label* metric that measures the correctness of class labels – it is implemented in the same way as the semantic segmentation task in the ScanNet

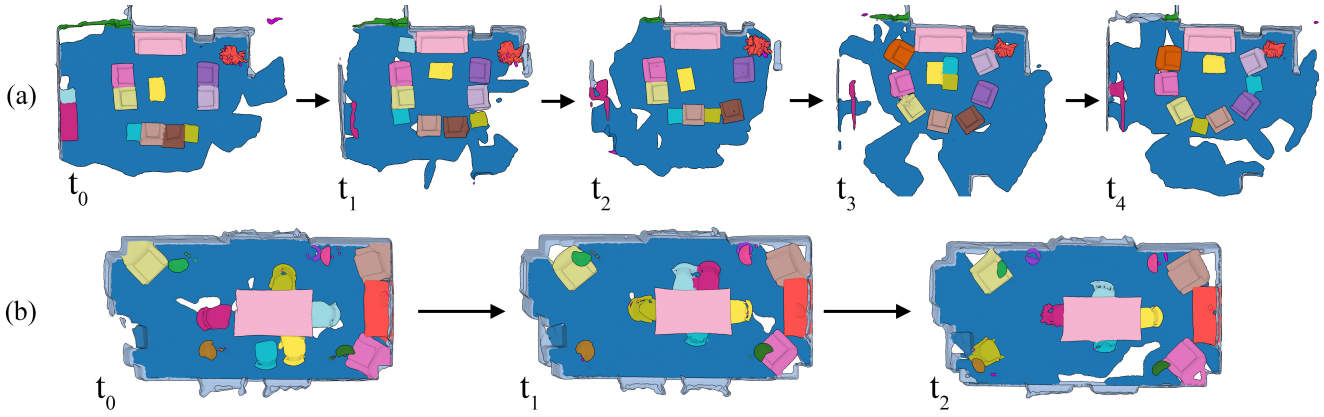



Figure 4: Inductive instance segmentation results. Given a segmentation at time $t_0$, our method is able to iteratively transfer instance labels to future times, even when the number of the objects in the scene changes.

Benchmark [10] and is reported as mean class IoU. The second is the *Semantic Instance* metric that measures the correctness of the object instance separations – it again comes from the ScanNet Benchmark [10] and is reported as mean Average Precision (IoU=0.5). Third, we propose a novel *Instance Transfer* metric, which specifically requires an agreement of instance indices across time. This metric is reported as mean IoU, where we count the number of points in both ground truth and prediction that share equivalent instance id. The *Instance Transfer* metric is much more challenging, as it requires associating objects with specific instance ids in different scans.

**Baseline:** Given the success of the recent deep models for the scene understanding (as shown on the leaderboard of [10]), it is interesting to compare the results of our algorithm to the best available method based on deep neural networks. One of the best available methods for 3D instance segmentation is MASC [33], which is based on semantic segmentation with SparseConvNet [21]. To test these methods on our tasks, we pre-trained the SparseConvNet and MASC models on ScanNet's training set. We performed fine-tuning of MASC with the ground-truth labels of first observation (time $t_0$) of each scene $S_0$ in our database. This fine-tuned model provides instance segmentation, which can be combined with the Hungarian method [30] to estimate instance associations across time. This sequence of steps provides a very strong baseline combining state-of-the-art methods for instance segmentation with an established algorithm for assignment.

### 4.1. Quantitative Results

**Evaluation and comparison:** Since we solve an inductive task (predict the answer at $t_i$, given an answer at $t_{i-1}$), it is not obvious how to initialize the system for our experiments. As our aim is to evaluate the inductive step alone, we chose to initialize time $t_0$ with a correct instance segmentation. That choice avoids confounding problems with de novo instance segmentation at $t_0$ with the main objective of the experiment. We have each algorithm in the experiment transfer the instance segmentation from $t_0$ to $t_1$, then transfer the result to $t_2$, and so on.

| Method | Semantic Label | Semantic Instance | Instance Transfer |
|---|---|---|---|
| SparseConvNet | 0.203 | - | - |
| MASC | 0.310 | 0.291 | 0.175 |
| MASC (fine-tuned) | 0.737 | 0.562 | 0.345 |
| Rescan | 0.859 | 0.837 | 0.650 |

Table 1: Comparison of our method to SparseConvNet [21] and MASC [33]. SparseConvNet does not produce instance labels, hence we omit reporting on the *Semantic Instance* and *Instance Transfer* task, and only fine-tune MASC.

We ran this experiment for our method in direct comparison to the baseline. Results for all three evaluation metrics are shown in Table 1. They show that our algorithm significantly outperforms competing methods. As expected, we see that the deep neural networks trained on the ScanNet training set [10] do not perform very well on our data without fine-tuning. After fine-tuning on the data in $S_0$, they do much better. Fine-tuning allows for a fair comparison, as both their and our methods have access to the same information from $S_0$ to predict labels for $S_i; i > 0$. Despite this, instance segmentation on later time steps still performs worse than our algorithm, and instance associations across time are poor. We attribute the difference to the fact that our method is instance-centric, where the segmentation is inferred from the estimated objects' arrangement. This is in stark opposition to methods like MASC, where the instances are inferred from a semantic segmentation.

**Ablation studies:** Second, we present the results of abla-

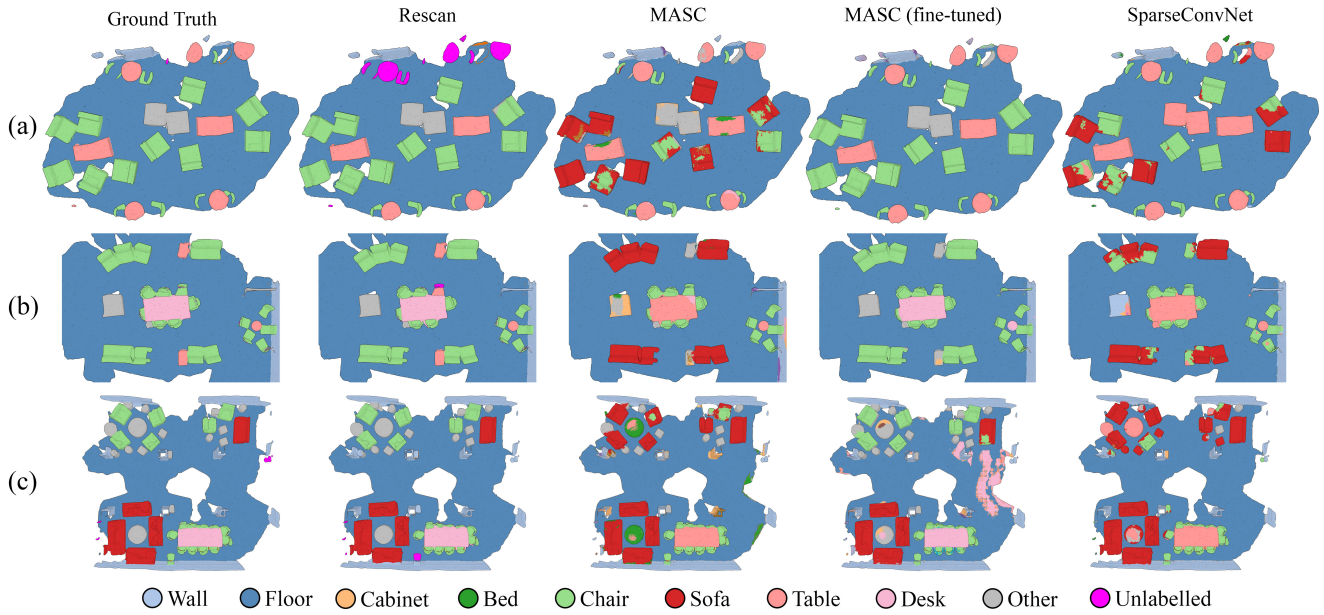



Figure 5: Qualitative comparison on the semantic segmentation task. Proposed method is able to provide high quality semantic labels as a result of instance segmentation transfer. Compared to competing methods, ours is able to produce better per object labels and does not confuse object classes.

tion studies that showcase the influence of various terms in our objective function on the results in a specific task. As seen in table 2, by far the most important term of our proposed objective is the *Coverage Term*. Without it, the objective function is discouraged from adding more objects. The optimization simply finishes with a single object added to the scene - as adding any more would lead to a decrease in other terms.

The second most important term, especially for the *Instance Transfer* task, is the *Hysteresis Term*. It is intuitive that lacking this term, the objective function is not encouraged to find an arrangement that will be consistent with previous object configurations. We note that when omitting this term, the semantic segmentation task achieves a slightly better result. The reason is that to prevent addition of superfluous objects the novel objects are assigned relatively low score (sec. 3.4.1). Without the *Hysteresis Term*, the proposed objective is free to insert additional objects - however their configuration is often not correct, leading to lower scores for other two tasks. This result suggests that there exists a better formulation of the hysteresis function - an interesting direction for future research.

The presence of the *Intersection Term* is important for the *Semantic Instance* and *Instance Transfer* tasks. Intuitively, the semantic segmentation score is unaffected as it is often the case that intersecting objects share the semantic class. The *Geometry Term* has the least influence on the results. This is not surprising, as the poses that survived the pose proposal stage (see sec. 3.3) were high scoring ones.

| Method | Semantic Label | Semantic Instance | Instance Transfer |
|---|---|---|---|
| No Coverage Term | 0.061 | 0.058 | 0.048 |
| No Geometry Term | 0.853 | 0.825 | 0.617 |
| No Intersection Term | 0.859 | 0.781 | 0.584 |
| No Hysteresis Term | 0.870 | 0.818 | 0.226 |
| Full Method | 0.859 | 0.837 | 0.650 |

Table 2: Ablation study showcasing the influence of objective function terms on each of the proposed tasks.

### 4.2. Qualitative Results

**Inductive segmentation transfer:** We showcase qualitative results for the *Instance Transfer* task using our method in figure 4. Again, in this task we use the ground-truth segmentation provided by the user at $t_0$ and transfer it to all other observations sequentially. The results of such segmentation transfer offer stable and well-localized instances. Even over multiple time-steps, our method is able to keep track of objects identities, providing us with information on their location and motion. Additionally, thanks to the fact that the objective function prefers minimal change, we are able to deal with challenging configurations. For example in 4a our method is able to correctly recover three coffee tables at time $t_3$, despite their proximity and visual similarity.

**Semantic segmentation:** Figure 5 showcases qualitative

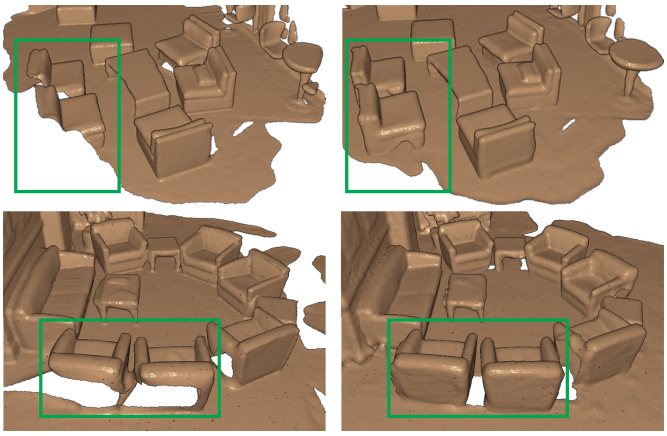


Figure 6: Model completion results. The left column shows two scans of a scene with moving objects. The right column shows our reconstruction of the scene using objects and locations from the temporal model $M$.

comparisons between our method and DNN-based methods [33, 21]. Without fine-tuning, the segmentation issues are obvious. Learned methods confuse labels like *sofa* and *chair*, which explains low scores in table 1. Fine-tuning helps reduce these effects - however we also see some overfitting errors. Our method is able to recover high quality semantic segmentation, where due to the fact that our approach is instance-centric, a single instance can not have more than a single semantic class. Our method's success is however dependent on the overlap between current and previous observations of $S$. When lots of novel objects appear, the *Hysteresis Term* might discouraging addition of all of them, as it aims to produce arrangement similar to previously observed ones (fig. 5a).

**Model completion results:** Our method for aggregating the observations of moving objects from multiple time steps allows it to produce more complete surface reconstructions than would be possible otherwise. Many other systems explicitly remove moving objects before creating a surface model (to avoid ghosting) [28]. Our approach uses the estimated object segmentations and transformations to aggregate points associated with each object $o_k$ to form a $G_k$ that is generally more complete than could be obtained from any one scan. Composing the aggregated $G_k$ using transformations $T_k$ in each object arrangement $A_i$ provides a model completion result (fig. 6).

**Failures:** We identify three main failure modes of our approach (fig. 7). The first issue arises due to the geometry focused nature of our approach. If the objects are only partially scanned, the pose proposal stage will not be able to recover highly scored poses. As such, these objects will simply not be added to the space of possible configurations that the optimization can choose from. The second is caused by the limited contribution of small objects to the scene

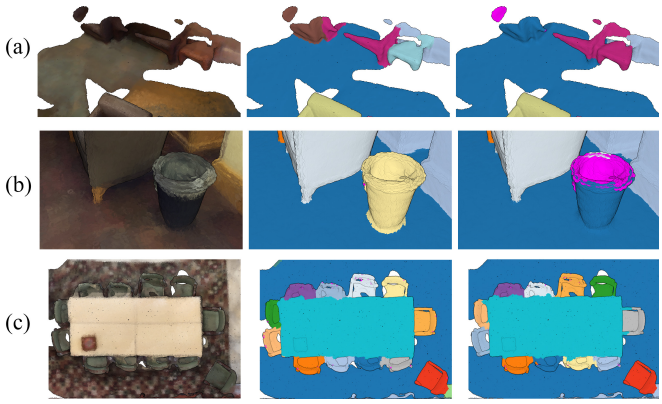



Figure 7: Failure modes of the proposed method. (a) Partial scanning prevents the pose proposal stage from generating plausible poses. (b) Small objects contribute little to the coverage term. If such objects undergo significant motion our algorithm might miss them. (c) When similar, partially scanned objects are considered, our method might not produce the correct permutation.

coverage score. Combined with a small *Hysteresis Term* value under significant motion, the objective function might prefer not adding these objects. Lastly, in cases like the one in figure 7c, an incorrect permutation of objects might have a higher objective value than the ground truth one. This effect is a combination of *Geometry Term* providing noisy scores for partial scans of visually similar objects (like the chairs around the table), and their relative spatial proximity, which makes the *Hysteresis Term* a poor discriminator.

## 5. Conclusion

This paper presents an algorithm for estimating the semantic instance segmentation for an RGBD scan of an indoor environment. The proposed algorithms is inductive – using a temporal scene model which subsumes previous observations, an instance segmentation of the novel observation is inferred and used to update the temporal model. Our experiments show better performance on a novel benchmark dataset in comparison to a strong baseline. Interesting directions for future work include inferring the segmentation at $t_0$, investigating RNN architectures (when larger datasets become available), and replacing terms of the objective function with learned alternatives.

## Acknowledgments

We would like to thank Angel X. Chang and Manolis Savva for insightful discussions. We also thank Graham et al. [21] and Liu et al. [33] for the comparison codes, and Dai et al. for the ScanNet data [10]. The project was partially supported by funding from the NSF (CRI 1729971 and VEC 1539014/1539099).

## References


[1] Rares Ambrus, Johan Ekekrantz, John Folkesson, and Patric Jensfelt. Unsupervised learning of spatial-temporal models of objects in a long-term autonomy scenario. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 5678–5685. IEEE, 2015. 2

[2] R. Ambru, N. Bore, J. Folkesson, and P. Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1854–1861, Sept 2014. 2

[3] Dragomir Anguelov, Rahul Biswas, Daphne Koller, Benson Limketkai, and Sebastian Thrun. Learning hierarchical object maps of non-stationary environments with mobile robots. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 10–17. Morgan Kaufmann Publishers Inc., 2002. 2

[4] T. Birdal and S. Ilic. Point pair features based object detection and pose estimation revisited. In *2015 International Conference on 3D Vision*, pages 527–535, Oct 2015. 3, 4

[5] Rahul Biswas, Benson Limketkai, Scott Sanner, and Sebastian Thrun. Towards object mapping in non-stationary environments with mobile robots. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, volume 1, pages 1014–1019. IEEE, 2002. 2

[6] Nils Bore, Johan Ekekrantz, Patric Jensfelt, and John Folkesson. Detection and tracking of general movable objects in large 3d maps. *arXiv preprint arXiv:1712.08409*, 2017. 2

[7] Frederic Bosche, Carl T Haas, and Burcu Akinci. Automated recognition of 3d cad objects in site laser scans for project 3d status visualization and performance control. *Journal of Computing in Civil Engineering*, 23(6):311–318, 2009. 3

[8] Angel 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. *International Conference on 3D Vision (3DV)*, 2017. 1

[9] M. Corsini, P. Cignoni, and R. Scopigno. Efficient and flexible sampling with blue noise properties of triangular meshes. *IEEE Transactions on Visualization and Computer Graphics*, 18(6):914–924, June 2012. 3

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 6, 8

[11] Andrew Delong, Anton Osokin, Hossam N. Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. *International Journal of Computer Vision*, 96(1):1–27, Jan 2012. 5

[12] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 998–1005, June 2010. 3

[13] Johan Ekekrantz, Nils Bore, Rares Ambrus, John Folkesson, and Patric Jensfelt. Towards an adaptive system for lifelong object modelling. *ICRA Workshop: AI for Long-term Autonomy*, 2016. 2

[14] J. Engel, T. Schöps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision (ECCV)*, September 2014. 3

[15] Thomas Fäulhammer, Rareş Ambruş, Chris Burbridge, Michael Zillich, John Folkesson, Nick Hawes, Patric Jensfelt, and Markus Vincze. Autonomous learning of object models on a mobile robot. *IEEE Robotics and Automation Letters*, 2(1):26–33, 2017. 2

[16] Marius Fehr, Fadri Furrer, Ivan Dryanovski, Jürgen Sturm, Igor Gilitschenski, Roland Siegwart, and Cesar Cadena. Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 5237–5244. IEEE, 2017. 2

[17] Ross Finman, Thomas Whelan, Liam Paull, and John J Leonard. Physical words for place recognition in dense rgb-d maps. In *ICRA workshop on visual place recognition in changing environments*, 2014. 2

[18] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. 4

[19] Garratt Gallagher, Siddhartha S Srinivasa, J Andrew Bagnell, and Dave Ferguson. Gatmo: A generalized approach to tracking movable objects. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 2043–2048. IEEE, 2009. 2

[20] Mani Golparvar-Fard, Feniosky Pena-Mora, and Silvio Savarese. Automated progress monitoring using unordered daily construction photographs and ifc-based building information models. *Journal of Computing in Civil Engineering*, 29(1):04014025, 2012. 3

[21] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. *CVPR*, 2018. 6, 8

[22] Kevin Han and Mani Golparvar-Fard. Bim-assisted structure-from-motion for analyzing and visualizing construction progress deviations through daily site images and bim. In *Congress on Computing in Civil Engineering, Proceedings*, volume 2015, pages 596–603, 06 2015. 3

[23] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, D. Hebesberger, T. Kortner, R. Ambrus, N. Bore, J. Folkesson, P. Jensfelt, L. Beyer, A. Hermans, B. Leibe, A. Aldoma, T. Faulhammer, M. Zillich, M. Vincze, E. Chinellato, M. Al-Omari, P. Duckworth, Y. Gatsoulis, D. C. Hogg, A. G. Cohn, C. Dondrup, J. Pulido Fentanes, T. Krajnik, J. M. Santos, T. Duckett, and M. Hanheide. The strands project: Long-term autonomy in everyday environments. *IEEE Robotics Automation Magazine*, 24(3):146–156, Sep. 2017. 2

[24] Evan Herbst, Peter Henry, and Dieter Fox. Toward online 3-d object segmentation and mapping. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 3193–3200. IEEE, 2014. 2

[25] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *International Conference on 3D Vision (3DV)*, 2016. 1

[26] Kevin Karsch, Mani Golparvar-Fard, and David Forsyth. Constructaide: analyzing and visualizing construction sites through photographs and building models. *ACM Transactions on Graphics (TOG)*, 33(6):176, 2014. 3

[27] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association. 5

[28] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In *Proceedings of Joint 3DIM/3DPVT Conference (3DV)*, page 8, 06 2013. 8

[29] Tomáš Krajník, Jaime P Fentanes, Joao M Santos, and Tom Duckett. Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Transactions on Robotics*, 33(4):964–977, 2017. 2

[30] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[31] Minhaeng Lee and Charless C. Fowlkes. Space-time localization and mapping. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3932–3941, 10 2017. 2

[32] Kok lim Low. Linear least-squares optimization for point-toplane icp surface registration. Technical report, University of North Carolina at Chapel Hill, 2004. 3

[33] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation, 2019. 6, 8

[34] Ricardo Martin-Brualla, David Gallup, and Steven M Seitz. Time-lapse mining from internet photos. *ACM Transactions on Graphics (TOG)*, 34(4):62, 2015. 3

[35] Kevin Matzen and Noah Snavely. Scene chronology. In *Proc. European Conf. on Computer Vision (ECCV)*, 2014. 3

[36] Richard Newcombe, Dieter Fox, and Steve Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2015. 2

[37] Danijel Rebolj, Zoran Pučko, Nenad Čuš Babič, Marko Bizjak, and Domen Mongus. Point cloud quality requirements for scan-vs-bim based automated construction progress monitoring. *Automation in Construction*, 84:323–334, 2017. 3

[38] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152, May 2001. 3

[39] João Machado Santos, Tomáš Krajník, and Tom Duckett. Spatio-temporal exploration strategies for long-term autonomy of mobile robots. *Robotics and Autonomous Systems*, 88:116–126, 2017. 2

[40] G. Schindler, F. Dellaert, and S. B. Kang. Inferring temporal order of images from 3d structure. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, June 2007. 3

[41] Dirk Schulz and Wolfram Burgard. Probabilistic state estimation of dynamic objects with a moving mobile robot. *Robotics and Autonomous Systems*, 34(2-3):107–115, 2001. 2

[42] Young Min Shin, Minsu Cho, and Kyoung Mu Lee. Multi-object reconstruction from dynamic scenes: An object-centered approach. *Comput. Vis. Image Underst.*, 117(11), Nov. 2013. 2

[43] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *Proceedings of the IEEE international conference on computer vision*, pages 233–240, 2013. 2

[44] Yelda Turkan, Frederic Bosche, Carl T Haas, and Ralph Haas. Automated progress tracking using 4d schedule and 3d sensing technologies. *Automation in Construction*, 22:414–421, 2012. 3

[45] Sebastian Tuttas, Alexander Braun, André Borrmann, and Uwe Stilla. Acquisition and consecutive registration of photogrammetric point clouds for construction progress monitoring using a 4d bim. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 85(1):3–15, 2017. 3

[46] Chieh-Chih Wang and Chuck Thorpe. Simultaneous localization and mapping with detection and tracking of moving objects. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 3, pages 2918–2924. IEEE, 2002. 2

[47] Chieh-Chih Wang, Charles Thorpe, and Sebastian Thrun. Online simultaneous localization and mapping with detection and tracking of moving objects: Theory and results from a ground vehicle in crowded urban areas. In *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, volume 1, pages 842–849. IEEE, 2003. 2

[48] Feilong Yan, Andrei Sharf, Wenzhen Lin, Hui Huang, and Baoquan Chen. Proactive 3d scanning of inaccessible parts. *ACM Trans. Graph.*, 33(4):157:1–157:8, July 2014. 2

[49] Jay Young, Valerio Basile, Markus Suchi, Lars Kunze, Nick Hawes, Markus Vincze, and Barbara Caputo. Making sense of indoor spaces using semantic web mining and situated robot perception. In *European Semantic Web Conference*, pages 299–313. Springer, 2017. 2