

# Seeing Behind Objects for 3D Multi-Object Tracking in RGB-D Sequences

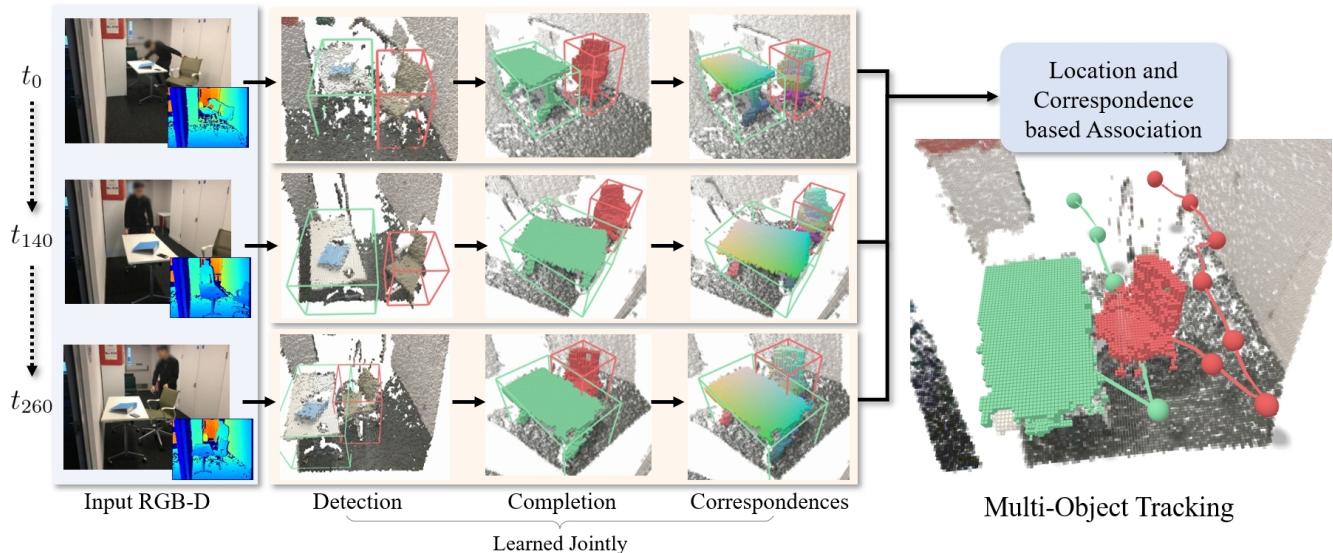
Norman Müller<sup>1</sup>Yu-Shiang Wong<sup>2</sup>Niloy J. Mitra<sup>2,3</sup>Angela Dai<sup>1</sup>Matthias Nießner<sup>1</sup><sup>1</sup>Technical University of Munich<sup>2</sup>University College London<sup>3</sup>Adobe Research

Figure 1. Our method learns to see behind objects in RGB-D sequences in order to achieve robust dynamic object tracking; we predict the complete underlying geometry of each object beyond the observed view, which enables finding correspondences which can more reliably persist over time, under various view changes and object motion. From an input RGB-D frame, we first perform 3D object detection, then jointly infer for each object its complete geometry and dense correspondence mapping to its canonical space. These correspondences on the predicted complete object geometry help to provide robust multi-object tracking over time.

## Abstract

*Multi-object tracking from RGB-D video sequences is a challenging problem due to the combination of changing viewpoints, motion, and occlusions over time. We observe that having the complete geometry of objects aids in their tracking, and thus propose to jointly infer the complete geometry of objects as well as track them, for rigidly moving objects over time. Our key insight is that inferring the complete geometry of the objects significantly helps in tracking. By hallucinating unseen regions of objects, we can obtain additional correspondences between the same instance, thus providing robust tracking even under strong change of appearance. From a sequence of RGB-D frames, we detect objects in each frame and learn to predict their complete object geometry as well as a dense correspondence mapping into a canonical space. This allows us to derive 6DoF poses for the objects in each frame, along with their corre-*

*spondence between frames, providing robust object tracking across the RGB-D sequence. Experiments on both synthetic and real-world RGB-D data demonstrate that we achieve state-of-the-art performance on dynamic object tracking. Furthermore, we show that our object completion significantly helps tracking, providing an improvement of 6.5% in mean MOTA.*

## 1. Introduction

Understanding how objects move over time is fundamental towards higher-level perception of real-world environments, with applications ranging from mixed reality to robotic perception. In the context of static scenes, significant progress has been made in RGB-D tracking and reconstruction [22, 17, 23, 32, 5, 9]; however, the assumption of a static environment significantly limits applicability to real-world environments which are often dynamic, with objects

moving over time. In the case of scenes where a number of objects might be rigidly moving, robust tracking remains a significant challenge, as views and occlusion patterns of the objects can change appreciably over time.

Several approaches have been developed to address the problem of dynamic object tracking in RGB-D sequences by detecting objects and then finding correspondences between frames [24, 25, 33]. While results have shown notable promise, these methods only consider the observed geometry of the objects, and so tracking objects under faster object or camera motion can result in insufficient overlap of observed geometry to find reliable correspondences, resulting in tracking failure.

To address these challenges, we observe that humans can effectively track objects by leveraging prior knowledge of the underlying object geometry, which helps to constrain the problem even under notable view changes or significant occlusions. Thus, our key idea is to learn to ‘see behind objects’ by *hallucinating the complete object geometry in order to aid object tracking*. We learn to jointly infer for each object its complete geometry as well dense tracking correspondences, providing 6DoF poses for the objects for each frame.

From an RGB-D sequence, we formulate an end-to-end approach to detect objects, characterized by their 3D bounding boxes, then predict for each object its complete geometry as well as a dense correspondence mapping to its canonical space. We then leverage a differentiable pose optimization based on the predicted correspondences of the complete object geometry to provide the object poses per frame as well as their correspondence within the frames.

Our experiments show that our joint object completion and tracking provides notably improved performance over state of the art by 6.5% in MOTA. Additionally, our approach provides encouraging results for scenarios with challenging occlusions. We believe this opens up significant potential for object-based understanding of real-world environments.

## 2. Related Work

**RGB-D Reconstruction of Static Scenes** Scanning and reconstruction 3D surfaces of static environments has been widely studied [22, 17, 5, 32, 9], with state-of-the-art reconstruction approaches providing robust camera tracking of large scale scenes. While these methods show impressive performance, they rely on a core, underlying assumption of a static environment, whereas an understanding of object movement over time can provide a profound, object-based perception.

Various approaches have also been developed for static scene reconstruction to simultaneously reconstruct the scene while also segmenting the observed geometry into semantic instances [28, 27, 20, 19]. Notably, Hou et

al. [15] propose to jointly detect objects as well as infer their complete geometry beyond the observed geometry, achieving improved instance segmentation performance; however, their method still focuses on static environments. In contrast, our approach exploits learning the complete object geometry in order to object tracking in dynamic scenes.

**RGB-D Object Tracking** Several approaches have been proposed towards understanding dynamic environments by object tracking. To achieve general non-rigid object tracking, research focuses on the single object scenario, typically leveraging as-rigid-as-possible registration [34, 21, 16, 10, 13, 4]. For multiple object tracking, object rigidity is assumed, and objects are detected and then tracked over time. In the context of SLAM, SLAMMOT [30], and CoSLAM [35] demonstrated detection and tracking of objects, operating with sparse reconstruction and tracking. Co-Fusion [24], MID-Fusion [33], and MaskFusion [25] demonstrated dense object tracking and reconstruction, with promising results for dynamic object tracking, but can still suffer noticeably from occlusions and view changes, as only observed geometry is considered. Our approach not only reconstructs the observed geometry of each object, but infers missing regions that have not been seen, which is crucial to achieve robust object tracking under these challenging scenarios.

## 3. Method Overview

Our method takes as input an RGB-D sequence, and learns to detect object instances, and for each instance the per-frame 6DoF poses and dense correspondences within the frames. We then associate the predicted locations and correspondences to obtain object tracking over time.

Each RGB-D frame of the sequence is represented by a sparse grid  $\mathcal{S}_i$  of surface voxels and a dense truncated signed distance field (TSDF)  $\mathcal{D}_i$ .

The TSDF for an RGB-D frame is obtained by back-projecting the observed depth values, following volumetric fusion [7].

As output, we characterize each detected object in every frame with a 3D occupancy mask representing its complete geometry along with a dense grid of correspondences to the object’s canonical space, from which we compute the 6DoF pose. We then use the complete correspondence prediction to associate objects across time steps, resulting in robust multi-object tracking over time.

From the input sparse surface grid, we detect objects by regressing their 3D object centers and extents, and cluster them into distinct bounding box proposals.

For each object proposal, we crop the TSDF volume using the respective bounding box, and use this information to predict the object’s complete geometry as a dense oc-

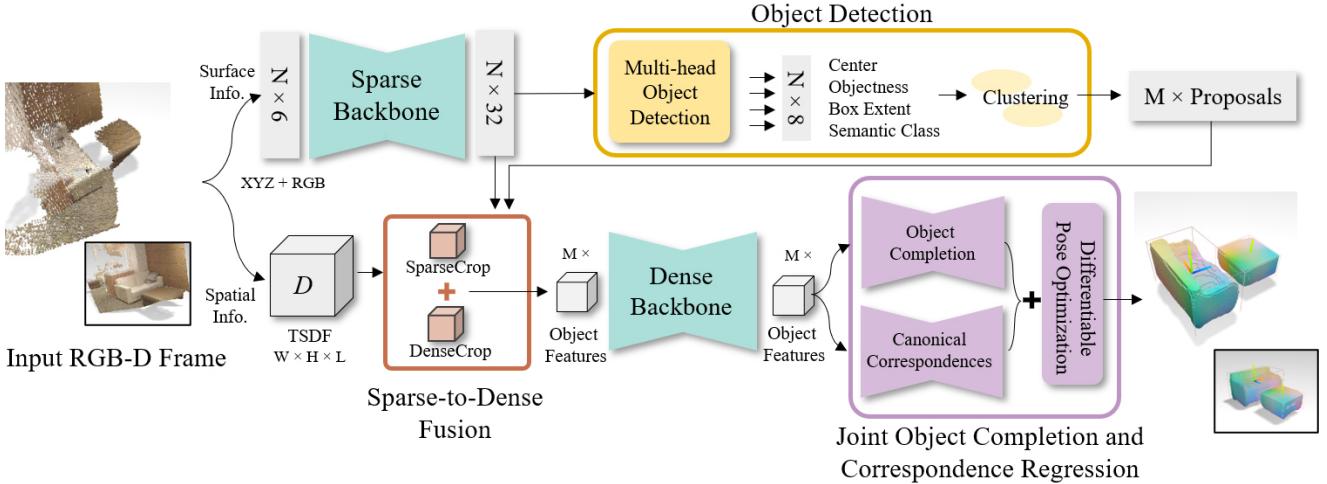


Figure 2. Overview of our network architecture for joint object completion and tracking. From a TSDF representation of an RGB-D frame, we employ a backbone of sparse 3D convolutions to extract features. We then detect objects characterized by 3D bounding boxes, and predict for each object both the complete object geometry beyond the view observation as well as dense correspondences in a canonical space; the correspondences on the complete geometry then inform a differentiable pose optimization to produce object pose estimates and within-frame dense correspondences. By predicting correspondences not only in observed regions but also unobserved areas, we can provide strong correspondence overlap under strong object or camera motion, enabling robust dynamic object tracking.

occupancy grid as well as its normalized object coordinates mapping the object to its canonical space.

We can then solve for the object pose using a differentiable Procrustes analysis.

To perform multi-object tracking across the RGB-D sequence, we associate instances across the frames based on 3D bounding box overlap as well as the 3D intersection-over-union of the predicted complete canonical geometry. Predicting the underlying geometric structure of each object enables our approach to maintain robustness under large camera pose changes or object movement, as we can associate the complete object geometry beyond the observed regions. Thus, from our object detection and then completion, we are able to find more correspondences which can persist over the full sequence of frames, providing more overlap for an object between frames, and resulting in more robust object instance tracking.

#### 4. Joint Object Completion and Tracking

From an RGB-D sequence, we first detect objects in each frame, then infer the complete geometry of each object along with its dense correspondences to its canonical space, followed by a differentiable pose optimization.

An overview of our network architecture for joint object completion and correspondence regression is shown in Figure 2. From an object detection backbone, we simultaneously predict an object’s complete geometry and dense correspondences, which informs its pose optimization. For a detailed architecture specification, we refer to the supplemental.

#### 4.1. Object Detection

We first detect objects from the sparse surface grid  $\mathcal{S}$  for each RGB-D frame by predicting their object bounding boxes. We extract features from the sparse surface grid using a series of sparse convolutions [12, 6] structured in encoder-decoder fashion, with features spatially bottlenecked to 1/16 of the original spatial resolution, and the output of the final decoder layer equal to the original spatial resolution. The feature map  $F$  from the last decoder layer is passed as input to a multi-head object detection module. The detection module predicts objectness, with each voxel  $v$  predicting  $O(v)$  as the score that  $v$  is associated with an object, the 3D center location  $C(v)$  of the object as a relative offset from  $v$ , and the 3D extents  $D(v)$  of the object as well as the semantic class  $S(v)$ . We then train using the following loss terms:

$$\begin{aligned} L_o &= BCE(O, O^t) \\ L_c &= \begin{cases} \frac{1}{2}(C - C^t)^2 & \text{for } |C - C^t| \leq 0.5, \\ |C - C^t| - \frac{1}{2}, & \text{otherwise} \end{cases} \\ L_d &= \begin{cases} \frac{1}{2}(D - D^t)^2 & \text{for } |D - D^t| \leq 0.5, \\ |D - D^t| - \frac{1}{2}, & \text{otherwise,} \end{cases} \\ L_s &= CE(S, S^t) \end{aligned}$$

with  $O^t$  denoting the target objectness as a binary mask of the target objects’ geometry, and  $C^t$ ,  $D^t$  and  $S^t$  the target object centers, extents and semantic class, respectively, defined within the mask of the target objects’ geometry.

To obtain the final object proposals, we perform a mean-shift clustering (20 steps, with 8 voxel radius) on the predicted center coordinates of the voxels which produce a positive objectness score. From the resulting instance clusters, we filter out small clusters of less than 50 elements. On the remaining clusters, we perform average pooling on the bounding box extent predictions and majority voting on the highest scoring semantic classes for final object location, shape and semantic class prediction.

**Sparse-to-Dense Fusion.** For each detected object and its predicted box, we then crop the corresponding sparse features  $f_k$  from  $F$  as well as the dense TSDF grid  $\mathcal{D}$ . We map the sparse cropped features densely and add the matching TSDF values over the feature channels to obtain  $f'_k$ . We can then leverage this feature to inform object completion and correspondence regression in both observed and unobserved space.

## 4.2. Object Completion

To predict the complete object geometry, we take the sparse-dense fused feature  $f'_k$  for an object  $k$ , which is then down-scaled by a factor of 2 using trilinear interpolation and passed through a series of dense 3D convolutions, structured in encoder-decoder fashion to obtain dense object features  $f_k^o$ . We then apply another series of dense 3D convolutional layers on  $f_k^o$  to predict the complete object geometry  $m_k$  as a binary mask trained by binary cross entropy with the target occupancy grid.

## 4.3. Object Correspondences

We predict for each object a dense correspondence mapping  $c_k$  to its canonical space, similar to the normalized object coordinate space of [31]. Using both  $c_k$  and the object geometry  $m_k$ , we can perform a robust pose optimization under the correspondences.

The correspondences  $c_k$  are predicted from the object feature map  $f_k^{o'}$  by a series of dense 3D convolutions structured analogously to the object geometry completion, outputting a grid of 3D coordinates in the canonical space of the object. We apply an  $l_1$  loss to the  $c_k$ , evaluated only where target object geometry exists.

To obtain the object pose in the frame, we take the correspondences from  $c_k$  where there is object geometry (using target geometry for training, and predicted geometry at test time), and optimize for the object rotation and scale under the correspondences using a differentiable Procrustes analysis.

We aim to find scale  $c^*$ , rotation  $R^*$  and translation  $t^*$  that bring together predicted object coordinates  $P_o$  with their predicted canonical representation  $P_n$ :

$$c^*, R^*, t^* := \underset{c \in \mathbb{R}^+, R \in SO_3, t \in \mathbb{R}^3}{\operatorname{argmin}} \|P_o - (cR \cdot P_n + t)\|. \quad (1)$$

With means  $\mu_i$  and variances  $\sigma_i$  of  $P_i$ ,  $i \in \{o, n\}$ , we perform a differentiable SVD of  $(P_o - \mu_o)(P_n - \mu_n)^T = UDV^T$ . According to [29], with  $S = \operatorname{diag}(1, 1, \det(UV^T))$ , we obtain the optima

$$c^* = \frac{1}{\sigma_n} \operatorname{tr}(DS), R^* = USV^T, \text{ and } t^* = \mu_o - c^* R^* \mu_n. \quad (2)$$

We employ a Frobenius norm loss on the estimated rotation matrix, an  $\ell_1$  loss on the predicted scale, and an  $\ell_2$  loss on the translation.

Since objects possessing symmetry can result in ambiguous target rotations, we take the minimum rotation error between the predicted rotation and the possible valid rotations based on the object symmetry.

## 4.4. Object Tracking

Finally, to achieve multi-object tracking over the full RGB-D sequence, we associate object proposals across time steps, based on location and canonical correspondences. Each detected object has a predicted bounding box and canonical object reconstruction, represented as a  $64^3$  grid by mapping the dense correspondences in the predicted object geometry to canonical space. To fuse detections over time into tracklets, we construct associations in a frame-by-frame fashion; we start with initial tracklets  $T^i$  for each detected object in the first frame.

Then, for each frame, we compute pairwise distances between current tracklets  $T^i$  and incoming proposals  $D^j$  based on the 3D IoU of their bounding boxes. We employ the Hungarian algorithm [18] to find the optimal assignment of proposals to tracklets, and reject any matches with 3D IoU below 0.3. Any new object detections with no matches form additional new tracklets. The canonical object reconstruction for a tracklet is then updated as a running average of the canonical reconstructions for each object detection in that tracklet; we use a 4:1 weighting for the running mean for all our experiments. After computing the tracklets and their canonical reconstructions from the frames in sequential order, we then aim to match any objects which might have not have been matched in the greedy sequential process (e.g., seen from a very different view, but able to match to the full reconstruction from many views). For all tracklets and all non-assigned proposals, we compute pairwise distances using a 3D volumetric IoU of the canonical representations (binarized at threshold 0.5). We again compute the optimal assignment and reject a matching if this mask IoU is below 0.3.

We find that by matching objects based on their canonical correspondences, we observe higher matching accuracy, leading to robust object tracking (see Section 5).

MOTA(%)	bathtub	bed	bookshelf	cabinet	chair	desk	sink	sofa	table	toilet	seq. avg
MaskFusion [25]	27.7	76.4	25.4	24.4	25.3	33.8	39.2	5.7	45.8	27.7	17.2
MID-Fusion [33]	<b>55.8</b>	<b>100</b>	<b>94.7</b>	21.7	38.6	45.8	63.9	9.6	53.8	35.7	30.1
F2F-MaskRCNN	25.7	<b>100</b>	73.7	15.2	28.3	<b>79.2</b>	<b>73.2</b>	<b>21.2</b>	59.6	33.9	35.8
Ours (no corr., no compl.)	39.8	54.5	22.6	21.8	27.2	37.5	49.5	13.8	60.4	36.7	29.3
Ours (no corr.)	39.8	54.5	24.0	23.2	32.2	37.5	50.3	13.8	61.8	38.1	30.6
Ours (no compl.)	24.9	45.5	50.0	26.1	42.3	66.4	63.3	18.0	63.2	38.0	35.6
Ours	24.9	45.5	50.1	<b>26.1</b>	<b>51.8</b>	66.4	63.3	17.3	<b>67.4</b>	<b>49.0</b>	<b>42.3</b>

Table 1. Evaluation of MOTA on DYNTHSYNTH. Our approach to jointly predict complete object geometry along with tracking provides robust correspondences over the full object rather than only the observed regions, resulting in notably improved tracking in comparison to our approach without object completion (*no compl.*), purely IoU based matching (*no corr.*) as well as state of the art.

#### 4.5. Training Details

We train our joint object completion and correspondence regression on a single Nvidia GeForce RTX 2080, using an ADAM optimizer with learning rate 0.001 and weight decay of 1e-5. We use a batch size of 2, and up to 10 proposals per input. To provide initial stable detection results, we first train the object detection backbone for 100K iterations, and then introduce the object completion and correspondence prediction along with the differentiable pose optimization, training the full model end-to-end for another 250K iterations until convergence. Full training takes approximately 72 hours.

We weight the object center and extent loss,  $L_c$  and  $L_d$  by 0.1, as they are evaluated in voxel units with have larger absolute value. After a warm-up phase of 100k iterations, where segmentation, detection and completion are trained individually, we weight the completion and correspondence loss by 4, and the rotation, translation and scale loss by 0.2, 0.1, 0.1, respectively, to bring the loss values into similar ranges.

### 5. Results

We evaluate our approach both quantitatively and qualitatively on synthetic RGB-D sequences of moving objects, as well as on real-world RGB-D data. We use a synthetic dataset, DYNTHSYNTH, which contains 3,300 RGB-D sequences of indoor scenes (2900/300/100 train/val/test), comprising 97,626 frames. We focus on detecting and tracking objects of 10 class categories covering a variety of bedroom, living room, and bathroom furniture. Each sequence contains camera trajectories and an object moving parallel to the ground, and ground truth object symmetries are provided.

As ground truth is available by nature of the synthetic data generation, we can train and fully evaluate our approach on DYNTHSYNTH. We also evaluate our object pose estimation on real-world, static RGB-D scans from the ScanNet data set [8] with ground truth object annotations provided by Scan2CAD [1]. We follow the offi-

cial train/val/test split with Scan2CAD annotations with 944/149/100 scans, resulting in 114,000 frames (sampled every 20th frame from the video sequences).

**Evaluation metrics.** To evaluate our dynamic object tracking, we adopt the Multiple Object Tracking Accuracy metric [2], which summarizes error from false positives, missed targets, and identity switches:

$$\text{MOTA} = 1 - \sum_t \frac{(m_t + fpt_t + mme_t)}{\sum_t gt} \quad (3)$$

where  $m_t$ ,  $fpt_t$ ,  $mme_t$  are number of misses, of false positives and of mismatches at time  $t$ .

A match is considered positive if its  $\ell_2$  distance to ground truth center is less than 25cm. The state-of-the-art approaches that we evaluate predict only surface correspondences, so we establish their trajectories by shifting from the initial pose towards the ground truth center. We report the mean MOTA over all test sequences.

**Comparison to state of the art.** In Table 1, we show that our approach to jointly complete and track objects provides significant improvement over state of the art on synthetic sequences from the DYNTHSYNTH dataset.

We compare to MaskFusion [25], a surfel-based approach for dense object tracking and reconstruction. MaskFusion’s segmentation refinement step is unable to handle objects with non-convex surface or disconnected topology due to the self-occlusion and its weighted surfel tracking mechanism is not robust in the highly dynamic scenes (i.e. new information tends to be discarded).

We evaluate against MID-Fusion [33], a volumetric octree-based, dense tracking approach; MID-Fusion use volumetric representation to alleviate the low recall issue of its detection backend. However, it has a limited ability to align occluded objects with the existed models and associate proposals under fast object movement such as the qualitative examples in Figure 3 and 4..

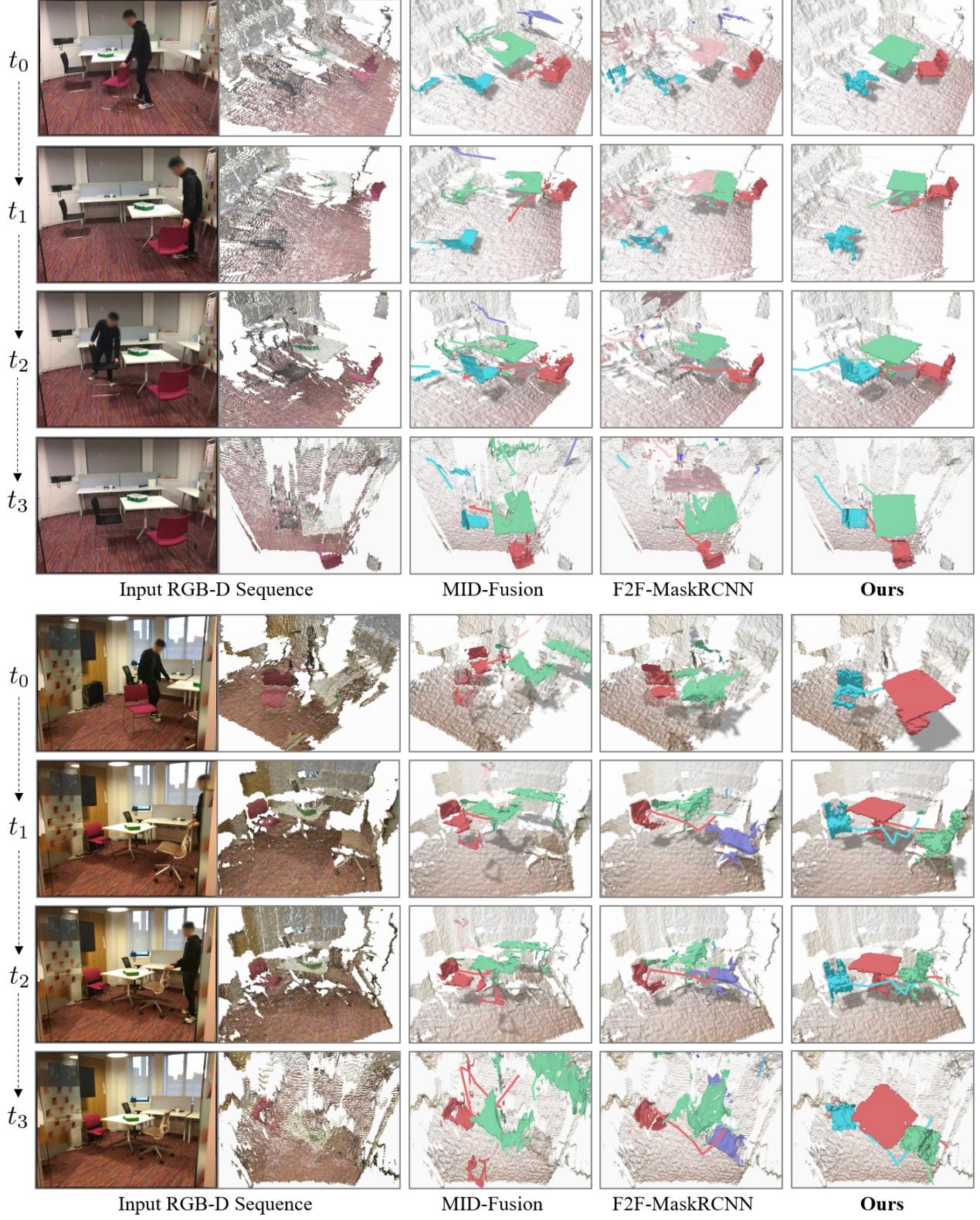


Figure 3. Our joint object completion and tracking on real-world RGB-D sequences maintains consistent objects tracks and accurate object shapes over time. The colors and the line segments show the instance ID and the estimated trajectories, respectively.

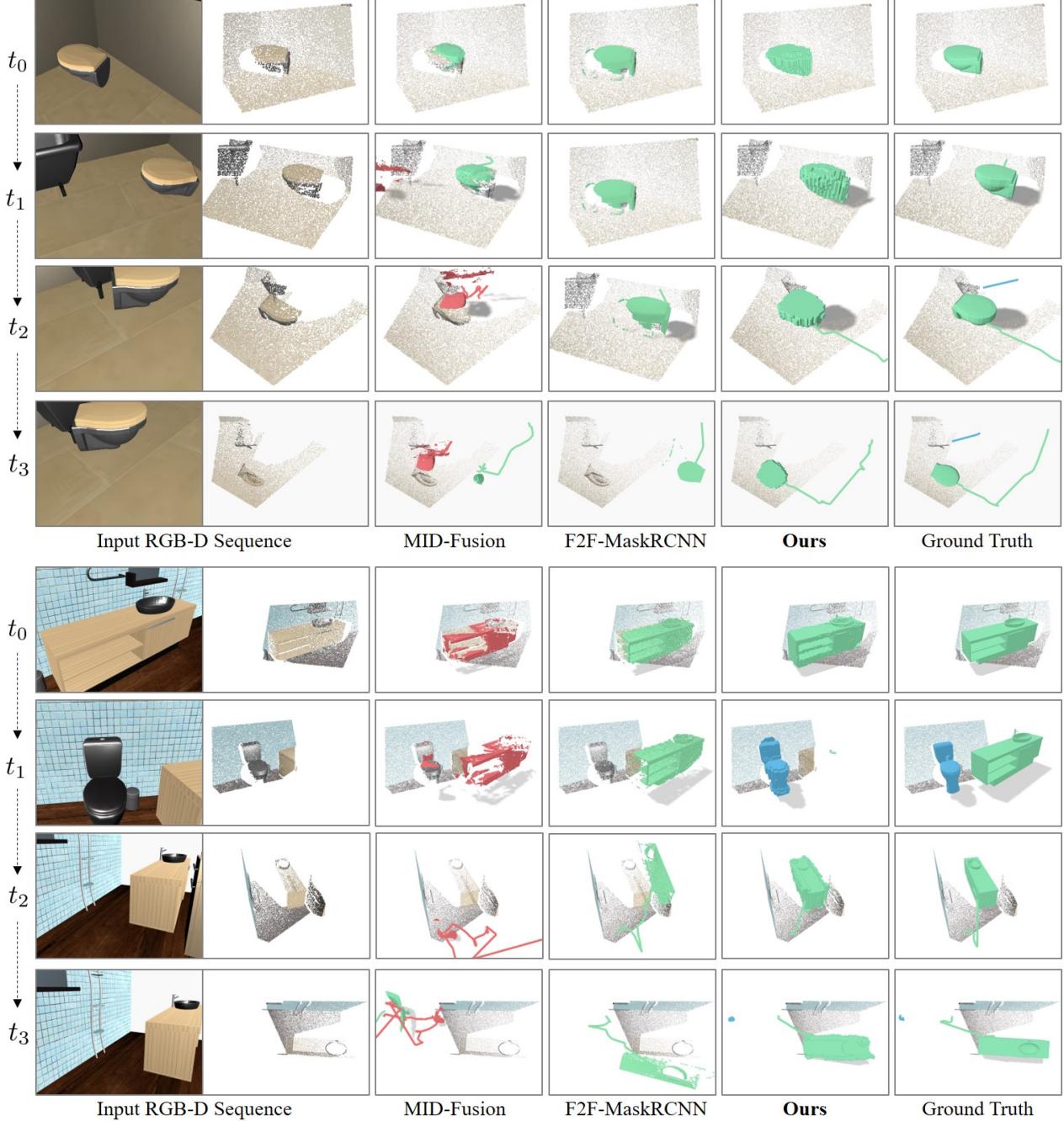


Figure 4. Qualitative comparison to state of the art on DYN SYNTH test sequences. Our approach predicting the complete object geometry maintains strong correspondence overlap even when objects or camera undergo stronger motions, resulting in notably more robust tracking than state-of-the-art approaches considering only the observed geometry.

Additionally, we provide a baseline approach which performs frame-to-frame tracking for each object using the Iterative Closest Point algorithm [3, 26], given 2D detection provided by Mask R-CNN [14] trained on DYN SYNTH (*F2F-MaskRCNN*). Searching correspondences between frames performs better under fast motion but it can-

not resolve the weak geometry signals issue [11] of the occluded objects such as the chair objects in Figure 3.

In contrast to these approaches which only reason based on the observed geometry from each view, our approach to infer the complete object geometry enables more robust and accurate object tracking.

**Does object completion help tracking?** We analyze the effect of our object completion on both dynamic object tracking performance as well as pose estimation in single frames. In Table 1, we evaluate our approach on variants without object completion (*no compl.*) or no correspondence-based object association (*no corr.*); When matching is fully based on 3D bounding box overlap, we notice a small improvement of tracking performance of the variant with completion (*no corr.*) over no completion (*no corr., no compl.*) of 1.6% mean MOTA. When association is based on canonical correspondences without using object completion (*no compl.*), we observe a performance gain of 5% mean MOTA. Utilizing object completion with canonical correspondences matching further improves the tracking performance by 6.7% mean MOTA and achieves best results (42.3% mean MOTA).

Additionally, we show that our joint object completion and tracking improves on pose estimation for each object in individual frames. Tables 2 and 3 evaluate our approach with and without object completion on RGB-D frames from synthetic DYN SYNTH data and real-world ScanNet [8] data, respectively. We similarly find that for object pose estimation, inferring the complete underlying geometric structure of the objects provides more accurate object pose estimation. Furthermore, we analyse in Figure 5 the tracking performance of our method with respect to the average completion performance on predicted tracklets. We observe that better completion also results in improved tracking, by facilitating correspondence in originally unobserved regions.

**Real-world dynamic RGB-D sequences.** In addition to the static RGB-D sequences of ScanNet [8], we apply our approach to eight real-world dynamic RGB-D sequences which we captured with a Structure Sensor<sup>1</sup> mounted to an iPad. In this scenario, we lack ground truth annotations, so we pre-train our model on DYN SYNTH and fine-tune on ScanNet+Scan2CAD data. Qualitative results are shown in Figure 3; our approach finds persistent correspondences on the predicted complete object geometry, enabling robust object pose estimation and surface tracking.

DynSynth	Med rot. err.	Med transl. err.
Ours (no compl.)	7.4°	15.4cm
Ours	<b>5.7°</b>	<b>12.3cm</b>

Table 2. Evaluation of object pose estimation on individual RGB-D frames from DYN SYNTH. Predicting the underlying geometry of each object enables more accurate object pose estimation in each frame.

ScanNet+Scan2CAD	Med rot. err.	Med transl. err.
Ours (no compl.)	16.6°	22.0cm
Ours	<b>13.3°</b>	<b>18.3cm</b>

Table 3. Evaluation of object pose estimation on individual RGB-D frames from ScanNet [8]. Understanding the complete object geometry enables more reliable correspondence prediction for object pose estimation.

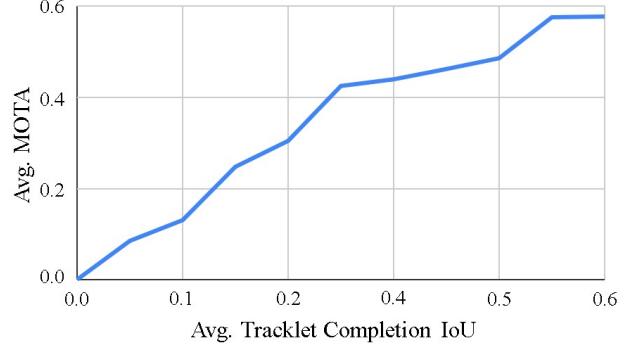


Figure 5. Average tracking performance against average completion performance evaluated on DYN SYNTH using our method. Better completion performance results in improved tracking, as correspondences can be more robustly established.

## 6. Conclusion

We introduce an approach for multi-object tracking in RGB-D sequences by learning to jointly infer the complete underlying geometric structure for each object as well as its dense correspondence mapping for pose estimation and tracking. By predicting object geometry in unobserved regions, we can obtain correspondences that are more reliably persist across a sequence, producing more robust and accurate object tracking under various camera changes and occlusion patterns. We believe that this provides significant promise in integration with a full reconstruction pipeline to perform live tracking and reconstruction of dynamic scenes towards object-based perception of environments.

## Acknowledgments

This work was supported by the ZD.B (Zentrum Digitalisierung.Bayern), a TUM-IAS Rudolf Mößbauer Fellowship, the ERC Starting Grant Scan2CAD (804724), and the German Research Foundation (DFG) Grant Making Machine Learning on Static and Dynamic 3D Data Practical. Yu-Shiang was partially supported by gifts from Adobe and Autodesk.

<sup>1</sup><https://structure.io/>

## References

- [1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Niessner. Scan2cad: Learning cad model alignment in rgb-d scans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5
- [2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5
- [3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 7
- [4] Aljaž Božič, Michael Zollhöfer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2020. 2
- [5] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5556–5565. IEEE, 2015. 1, 2
- [6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 3
- [7] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 2
- [8] 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. 5, 8
- [9] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3):24:1–24:18, 2017. 1, 2
- [10] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escalano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 2
- [11] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy. Geometrically stable sampling for the icp algorithm. In *Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings.*, pages 260–267, 2003. 7
- [12] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 3
- [13] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 2
- [14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7
- [15] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2020. 2
- [16] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pages 362–379. Springer, 2016. 2
- [17] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard A. Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew J. Davison, and Andrew W. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16–19, 2011*, pages 559–568, 2011. 1, 2
- [18] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 4
- [19] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41. IEEE, 2018. 2
- [20] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. 2
- [21] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2
- [22] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011, Basel, Switzerland, October 26–29, 2011*, pages 127–136, 2011. 1, 2
- [23] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 2013. 1
- [24] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, 2017. 2
- [25] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Sym-*

- posium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE, 2018. 2, 5
- [26] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pages 145–152. IEEE, 2001. 7
- [27] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013. 2
- [28] Keisuke Tateno, Federico Tombari, and Nassir Navab. When 2.5 d is not enough: Simultaneous reconstruction, segmentation and recognition on dense slam. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 2295–2302. IEEE, 2016. 2
- [29] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 4
- [30] Chieh-Chih Wang, Charles Thorpe, Sebastian Thrun, Martial Hebert, and Hugh Durrant-Whyte. Simultaneous localization, mapping and moving object tracking. *The International Journal of Robotics Research*, 26(9):889–916, 2007. 2
- [31] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4
- [32] Thomas Whelan, Stefan Leutenegger, Renato F. Salas-Moreno, Ben Glocker, and Andrew J. Davison. Elasticfusion: Dense SLAM without A pose graph. In *Robotics: Science and Systems XI, Sapienza University of Rome, Rome, Italy, July 13-17, 2015*, 2015. 1, 2
- [33] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5231–5237. IEEE, 2019. 2, 5
- [34] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4):1–12, 2014. 2
- [35] Danping Zou and Ping Tan. Coslam: Collaborative visual slam in dynamic environments. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):354–366, 2012. 2

## Appendix

In this appendix, we provide further details about our proposed method. Specifically, we describe the network architectures in detail in Section B and provide more quantitative results in Section A.

### A. Additional Quantitative Evaluation

We provide per-frame model performance on real-world ScanNet+Scan2CAD and the synthetic dataset DYN-SYNTH. In Table 4, we show class-wise detection results evaluated as mean average precision at a 3D IoU of 0.5 (mAP@0.5). The per-frame completion performance is evaluated in Table 5 using a mean average precision metric with mesh IoU threshold of 0.25 (mAP@0.25).

### B. Network Details

We detail the architecture of our network in Figure 6. We provide the convolution parameters as (n\_in, n\_out, kernel\_size, stride, padding), where stride and padding default to 1 and 0, respectively. Each convolution (except the last) is followed by batch normalization and a ReLU.

	bathtub	bed	bookshelf	cabinet	chair	desk	sink	sofa	table	toilet	mAP
DYN-SYNTH	49.3	38.4	12.5	6.3	44.1	46.8	27.6	32.3	38.4	63.1	35.8
ScanNet+Scan2CAD	38.7	-	12.9	4.6	41.2	-	-	26.4	29.2	-	25.6

Table 4. 3D Detection results on DYN-SYNTH and ScanNet with Scan2CAD targets at mAP@0.5.

	bathtub	bed	bookshelf	cabinet	chair	desk	sink	sofa	table	toilet	mAP
DYN-SYNTH	34.8	23.6	12.7	11.4	38.4	34.1	32.2	41.1	29.9	52.6	31.1
ScanNet+Scan2CAD	20.4	-	8.6	12.7	24.4	-	-	23.9	12.2	-	17.1

Table 5. Instance Completion results on DYN-SYNTH and ScanNet with Scan2CAD targets at mAP@0.25.

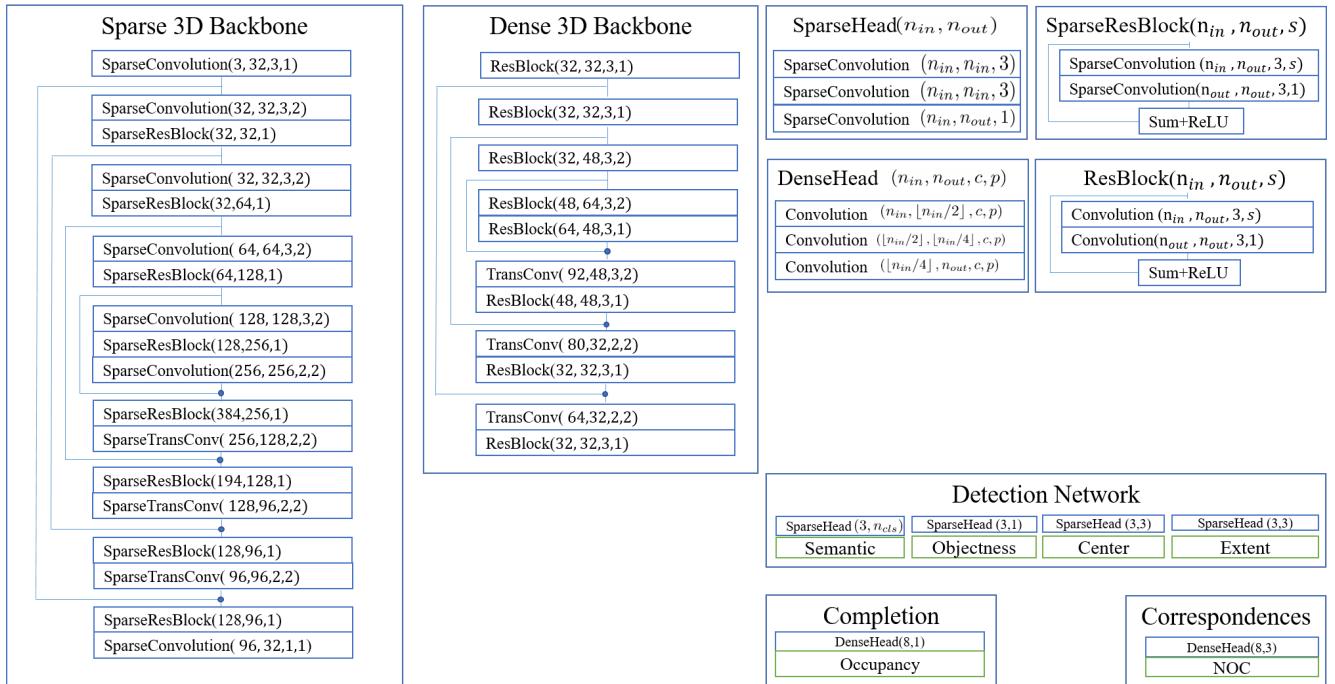


Figure 6. Network architecture specification for our approach. Dots indicate concatenation, outputs are highlighted in green.